Statistical Methodology 15 (2013) 115–120

Contents lists available at SciVerse ScienceDirect

Statistical Methodology

journal homepage: www.elsevier.com/locate/stamet

A confidence-set approach to signal denoising

Boris Ryabko [a,*], Daniil Ryabko [b]

[a] *Siberian State University of Telecommunications and Information Sciences, Institute of Computational Technology of Siberian Branch of Russian Academy of Science, Novosibirsk, Russia*

[b] *INRIA Lille, France*

ARTICLE INFO

Article history:
Received 30 September 2012
Received in revised form
18 March 2013
Accepted 18 May 2013

Keywords:
Time series
Confidence set
Filtering

ABSTRACT

The problem of filtering of finite-alphabet stationary ergodic time series is considered. A method for constructing a confidence set for the (unknown) signal is proposed, such that the resulting set has the following properties. First, it includes the unknown signal with probability γ, where γ is a parameter supplied to the filter. Second, the size of the confidence sets grows exponentially with a rate that is asymptotically equal to the conditional entropy of the signal given the data. Moreover, it is shown that this rate is optimal. We also show that the described construction of the confidence set can be applied to the case where the signal is corrupted by an erasure channel with unknown statistics.

1. Introduction

The problem of estimating a discrete signal $X_1, \ldots, X_t$ from a noisy version $Z_1, \ldots, Z_t$ has attracted the attention of many researchers due to its great importance for statistics, computer science, image processing, astronomy, biology, cryptography, information theory and many other fields. The main attention is usually focused on developing methods of estimation (denoising, or filtering) of the unknown signal, with the performance measured under a given fidelity criterion; see, for example, [7,8] and references therein. Such an approach can be related to the problem of a point estimation in statistics.

An alternative approach, often considered in mathematical statistics, is that of constructing confidence sets. That is, one tries to use the data to construct a set that includes the unknown parameter (in our case, the signal) with a prescribed probability, while trying to keep the size of the set as small

* Corresponding author. Tel.: +7 9139219968.
E-mail addresses: boris@ryabko.net, rbya@mail.ru (B. Ryabko), daniil@ryabko.net (D. Ryabko).

1572-3127/$ – see front matter
http://dx.doi.org/10.1016/j.stamet.2013.05.003

as possible (see, e.g. [5] for some classical examples). Such a set is usually constructed as the set of most likely values of the parameter.

In statistics, the two approaches can be considered complementary. However, in filtering, while the point-estimation approaches abound, the confidence-set counterpart is missing. Note that, in the presence of noise, the exact recovery of the signal could be impossible, and thus the point estimate is necessarily imperfect. The choice of a particular estimate of the signal out of many likely alternatives could be largely arbitrary. Moreover, the optimal choice may depend on the specific application involved. In such cases, a confidence-set estimate provides additional information that can potentially be used to select a more appropriate (for each given application) point estimate.

This is the approach and the problems considered in this work. We consider a model in which the underlying noiseless signal and the resulting corrupted (noisy) signal (and thus the channel) are assumed to be stationary ergodic processes with finite alphabets. We mainly concentrate on the case where the probability distributions of the noiseless signal and the noisy channel are known. (Obviously, in such a case the distribution of the corrupted signal is known, too.) Besides, the case of an erasure channel with unknown distribution is also considered, because in this case the conditional distribution of the noiseless signal is known, even though the distribution of the noise is unknown. This gives a possibility to apply directly the methods proposed for the case of known statistics of the noiseless signal and noise. The results that we obtain establish the optimal rate of growth (with respect to time, or to the length of the signal) of the size of the confidence set, as well as a method for constructing such a set. The optimal rate turns out to be equal to the entropy of the signal given its noisy version.

The following examples illustrates the fact that a confidence set can provide additional information; the example also exposes the notation used further in the text. Suppose that the noiseless sequence is a text in English, corrupted by an erasure noise in such a way that the probability of each symbol to be erased does not depend on the symbol. Suppose that the given sequence is as follows:

$$Z_1, \dots, Z_{10} = \textit{Great fea} * .$$

In this example, we do not know the probability distribution of the original words. Instead (similarly to a method proposed in [13]) we will use the estimates obtained via a search engine (here we used `Google`) that gives frequencies of occurrence of a search term in a vast corpus of documents. The precision of these estimates is questionable (in particular, the outputs of a search engine vary to a certain extent), but it suffices for the purpose of this illustration. Using this method, we obtain the following probabilities:

$$\begin{aligned}
&X_1, \dots, X_{10} = \textit{Great fear}, \qquad P(X_1, \dots, X_{10}|Z_1, \dots, Z_{10}) = 0.664,\\
&X_1, \dots, X_{10} = \textit{Great feat}, \qquad P(X_1, \dots, X_{10}|Z_1, \dots, Z_{10}) = 0.335,\\
&\text{all other values}, \quad P(X_1, \dots, X_{10}|Z_1, \dots, Z_{10}) < 0.001.
\end{aligned}$$

We can see that using the point estimate one obtains only the first version (*Great fear*). Using the confidence set with γ, say, 0.99, one is able to see two possible texts (*Great fear* and *Great feat*), which have a very different meaning. It appears that the choice of the answer in this case may depend on the context or the application used.

The goal of this paper is to describe a construction of confidence sets and to give an estimate of their size, for the case when the signal and noise are stationary ergodic processes with finite alphabets. It is shown that for any $\gamma \in (0, 1)$ the size of the confidence set grows exponentially with the rate $h(X|Z)$, where $h(X|Z)$ is the limit (conditional) Shannon entropy. This result is valid for the case when the probability distributions of noiseless signal and noise are known, as well as for the case when the probability distribution of the signal is known and the noise is described by a stationary erasure channel with memory whose probability distribution is unknown. Moreover, we prove that the rate $h(X|Z)$ is minimal, which means that the suggested method of constructing confidence sets is asymptotically optimal.

It is worth noting that the information theory is deeply connected with statistics of time series and signal processing; see, for example, [3,6,10,9,12] and [7,8], correspondingly. In this paper a new

connection of this kind is established: it is shown that the Shannon entropy determines the rate of growth of the size of the confidence set for the signal, given its version corrupted by stationary noise.

2. Preliminaries

We consider the case where the signal $X = X_1, X_2, \dots$ and its noisy version $Z = Z_1, Z_2, \dots$ are described by stationary ergodic processes with finite alphabets $\mathbf{X}$ and $\mathbf{Z}$ respectively. (There may be arbitrary long-range dependencies between the variables.) It is assumed that probability distributions of both processes are known, and, hence, the statistical structure of the noise corrupting the signal $X = X_1, X_2, \dots$ is known, too. Introduce the short-hand notation $X_{1..t}$ for $X_1, \dots, X_t$, and analogously for Z.

The n-order Shannon entropy and the limit Shannon entropy are defined as follows:

$$h_n(X) = -\frac{1}{n+1} \sum_{u \in A^{n+1}} P_X(u) \log P_X(u), \qquad h(X) = \lim_{n\to\infty} h_n(X) \tag{1}$$

where $n \geq 0$, $P_X(u)$ is the probability that $X_1X_2 \dots X_{|u|} = u$ (this limit always exists, see, for example, [2,4]). Introduce also the conditional Shannon entropy

$$h_n(X|Z) = h_n(X, Z) - h_n(Z), \qquad h(X|Z) = \lim_{n\to\infty} h_n(X|Z). \tag{2}$$

The Shannon–McMillan–Breiman theorem for conditional entropies can be stated as follows.

Theorem 1 (*Shannon–McMillan–Breiman*). *$\forall \varepsilon > 0, \forall \delta > 0$, for almost all $Z_1, Z_2, \dots$ there exists n' such that if $n > n'$ then*

$$P\left\{\left|-\frac{1}{n}\log P(X_{1..n}|Z_{1..n}) - h(X|Z)\right| < \varepsilon\right\} \geq 1 - \delta. \tag{3}$$

The proof can be found in [1,2,4].

3. Confidence sets and their properties

Informally, for any $\gamma \in (0, 1)$ and any sequence $Z_1, \dots, Z_t$ we define the confidence set $\Psi^t_\gamma(Z_1, Z_2, \dots, Z_t)$ as follows: the set contains sequences $x_1, x_2, \dots, x_t$ whose probabilities $P(x_{1..t}|Z_{1..t})$ are maximal and sum to γ. This definition is not precise, since it is possible that the sum cannot be made equal to γ exactly. That is why the formal definition of the confidence set will use randomization.

For this purpose, we order all sequences $X_{1..t}$ according to their conditional probabilities, in decreasing order. That is, we enumerate all sequences $x_{1..t} \in \mathbf{X}^n$ in such a way that $(a_{1..t}) \in \mathbf{X}^t$ has a smaller index than $(b_{1..t}) \in \mathbf{X}^t$ if either $P(a_{1..t}|Z_{1..t}) > P(b_{1..t}|Z_{1..t})$, or $P(a_{1..t}|Z_{1..t}) = P(b_{1..t}|Z_{1..t})$ and $(a_{1..t})$ is lexicographically less than $(b_{1..t})$. Let j be the integer for which $\sum_{i=1}^{j-1} P(x^i_{1..t}|Z_{1..t}) \leq \gamma$ and $\sum_{i=1}^{j} P(x^i_{1..t}|Z_{1..t}) > \gamma$. If $\sum_{i=1}^{j-1} P(x^i_{1..t}|Z_{1..t}) = \gamma$, then define $\Psi^t_\gamma(Z_{1..t})$ as the set $\{x^1_{1..t}, \dots, x^{j-1}_{1..t}\}$. Otherwise, $\Psi^t_\gamma(Z_{1..t})$ also contains $j-1$ first elements, and additionally the element $x^j_{1..t}$ with probability $(\gamma - \sum_{i=1}^{j-1} P(x^i_{1..t}|Z_{1..t}))/P(x^j_{1..t}|Z_{1..t})$. (Note that this procedure is commonly used in mathematical statistics for making the confidence level exactly γ.) When talking about the sizes of the confidence sets we refer to their expected (with respect to the randomization) size.

Next, we estimate the size of the described confidence set.

Theorem 2. *Let an (unknown) signal $X = X_1X_2, \dots$ and its noisy version $Z = Z_1Z_2, \dots$ be stationary ergodic processes with finite alphabets. Then, for every $\gamma \in (0, 1)$, all $t \in \mathbb{N}$ and almost every $Z_1, \dots, Z_t$ the confidence set $\Psi^t_\gamma(Z_1, \dots, Z_t)$ contains the unknown $(X_1, \dots, X_t)$ with probability γ:*

$$P\{X_{1..t} \in \Psi^t_\gamma(Z_{1..t})\} = \gamma, \tag{4}$$

while, with probability 1, *the size of the set* $\Psi_\gamma^t(Z_1,\dots,Z_t)$ *grows exponentially with the exponent rate that is equal to the conditional entropy:*

$$\lim_{t\to\infty}\frac{1}{t}\log\mathbf{E}|\Psi_\gamma^t(Z_1,\dots,Z_t)| = h(X|Z) \quad a.s., \tag{5}$$

where the expectation is with respect to the randomization used in constructing the confidence sets.

Besides, the size of $\Psi_\gamma^t(Z_1,\dots,Z_t)$ *is asymptotically minimal. More precisely, let* $\Phi_\gamma^t(Z_{1..t})$ *be any confidence sets satisfying* $P\left(X_{1..t}\in\Phi_\gamma^t(Z_{1..t})\right)\geq\gamma$ *for almost all* $Z_{12\dots}$ *and for all* $t\in\mathbb{N}$. *Then, with probability* 1,

$$\liminf_{t\to\infty}\frac{1}{t}\log|\Phi_\gamma^t(Z_1,\dots,Z_t)|\geq h(X|Z). \tag{6}$$

Proof. The proof of (4) immediately follows from the construction of the set $\Psi_\gamma^t(Z_1Z_2\dots Z_t)$. The proof of (5) is based on (3). Take any $\varepsilon>0$ and any $\delta>0$ such that

$$1-\delta\geq\gamma. \tag{7}$$

From (3) we conclude that for almost all $Z_1,Z_2,\dots$ there exists n' such that (3) is valid if $n>n'$. Take any such n and rewrite (3) as follows:

$$P\left\{2^{-n(h(X|Z)+\varepsilon)}\leq P(X_{1..n}|Z_{1..n})\leq 2^{-n(h(X|Z)-\varepsilon)}\right\}\geq 1-\delta. \tag{8}$$

Thus, the probability of all strings $x_1,\dots,x_n$ for which we have $P(x_{1..n}|Z_{1..n})\geq 2^{-n(h(X|Z)+\varepsilon)}$ is at least $(1-\delta)$. Taking into account (7), we have

$$|\Psi_\gamma^t(Z_{1..n})|\leq\gamma/2^{-n(h(X|Z)+\varepsilon)},$$

so that

$$\frac{1}{n}\log|\Psi_\gamma^t(Z_{1..n})|\leq h(X|Z)+\varepsilon+O(1/n) \tag{9}$$

for $n>n'$. Having taken into account that (9) holds for every $\varepsilon>0$ we obtain that a.s.

$$\limsup_{t\to\infty}\frac{1}{t}\log\mathbf{E}|\Psi_\gamma^t(Z_1,\dots,Z_t)|\leq h(X|Z). \tag{10}$$

The opposite inequality (6) will be proven for the size of any confidence sets, implying (5).

To prove (6), we take, as before, any $\varepsilon>0$ and fix $\delta:=\gamma/2$. Then from some n on we have (8). Let Υ be a confidence set for this n and a certain γ. Define

$$\Phi=\left\{x_{1..n}: 2^{-n(h(X|Z)+\varepsilon)}\leq P(x_{1..n}|Z_{1..n})\leq 2^{-n(h(X|Z)-\varepsilon)}\right\}. \tag{11}$$

By definition, $\sum_{x_{1..n}\in\Upsilon}P(x_{1..n}|Z_{1..n})\geq\gamma$. From this and (8) we obtain

$$\sum_{x_{1..n}\in\Upsilon\cap\Phi}P(x_{1..n}|Z_{1..n})\geq\gamma-\delta.$$

From this and (11) we get

$$|\Upsilon|\geq|\Upsilon\cap\Phi|\geq(\gamma-\delta)2^{n(h(X|Z)-\varepsilon)}.$$

Hence,

$$\liminf_{t\to\infty}\frac{1}{n}\log|\Upsilon|\geq h(X|Z)-\varepsilon.$$

Since this inequality is true for any confidence set Υ and any $\varepsilon>0$, we obtain (6). Taking into account that (6) is true for $\Psi_\gamma^t(Z_1,\dots,Z_t)$, too, we obtain from (10), Eq. (5). □

4. Erasure channel with unknown statistics

In this section we consider the case when the channel statistics is unknown, but the channel has a specific form: it is an erasure channel, such that the probability of each symbol to be erased is the same for all symbols. We show that the confidence sets described above are asymptotically optimal in this case, too. The reason why this extension holds is that in this case the conditional probabilities $P(X_{1..n}|Z_{1..n})$ are known.

The formal description of the considered model is as follows. We still assume that there is a known stationary ergodic source generating the signal $X_1, X_2, \ldots$. The erasure channel is defined in the two following steps: first, there is a stationary ergodic process Θ generating letters from the alphabet $\{\Lambda, *\}$ and, second, the noisy channel is determined by the following "summation" of the (uncorrupted) sequence $X_1, X_2, \ldots$ and the noise sequence $\Theta_1, \Theta_2, \ldots$:

$$Z_i = \begin{cases} X_i & \text{if } \Theta_i = \Lambda \\ * & \text{if } \Theta_i = *. \end{cases}$$

Theorem 3. *Let an (unknown) signal $X = X_1X_2, \ldots$ and $Z_1, Z_2, \ldots$ be a stationary ergodic signal and its version corrupted by an unknown stationary erasure channel. Then, for every $\gamma \in (0, 1)$, all $t \in \mathbb{N}$ and almost every $Z_1, \ldots, Z_t$ the (above described) confidence set $\Psi^t_\gamma(Z_1, \ldots, Z_t)$ contains the unknown $(X_1, \ldots, X_t)$ with probability γ:*

$$P\{X_{1..t} \in \Psi^t_\gamma(Z_{1..t})\} = \gamma, \tag{12}$$

while, with probability 1, the size of the set $\Psi^t_\gamma(Z_1, \ldots, Z_t)$ grows exponentially with the exponent rate that is equal to the conditional entropy:

$$\lim_{t\to\infty} \frac{1}{t} \log \mathbf{E}|\Psi^t_\gamma(Z_1, \ldots, Z_t)| = h(X|Z) \quad a.s., \tag{13}$$

where the expectation is with respect to the randomization used in constructing the confidence sets.

Proof. It is enough to notice that, although the erasure channel is not known, the probabilities $P(X_{1..n}|Z_{1..n})$ are known. Therefore, the proof of this theorem is identical to that of Theorem 2. □

5. Discussion

To the best of our knowledge, the problem of constructing a confidence set for the unknown signal was not considered before, which is why there are many quite natural and obvious extensions and generalizations of the present work. First, it is interesting to consider this problem for certain specific classes of distributions of the signal and noise, such as i.i.d. and Markov sources. For these classes of sources it should be possible to obtain rates of convergence in those statements that in this work are only asymptotic, for example in (5).

Second, a natural question is to find a construction of the confidence set for the cases where the signal is multi-dimensional. This is particularly important for applications, many of which are concerned with denoising such objects as photographs or video fragments. Another interesting generalization is the case where the alphabets are (subsets of), for example, the Euclidean space. This generalization can be also interesting from the practical point of view. Finally, the case where statistics of the noise and/or signal are unknown is obviously of great theoretical and practical interest.

Acknowledgment

Some of these results were reported by the authors at ISIT 2011 see [11]. This research has been partially supported by the Russian Foundation of Basic Research, grant no. 12-07-00125 (first author), the French Ministry of Higher Education and Research, Nord-Pas-de-Calais Regional Council and FEDER (CPER 2007–2013), and ANR project Lampada ANR-09-EMER-007 (second author).

References

[1] P. Algoet, T. Cover, A sandwich proof of the Shannon–McMillan–Breiman theorem, Annals of Probability 16 (1988) 899–909.
[2] T.M. Cover, J.A. Thomas, Elements of Information Theory, Wiley-Interscience, New York, NY, USA, 2006.
[3] I. Csiszar, P.C. Shields, Notes on information theory and statistics, in: Foundations and Trends in Communications and Information Theory, 2004.
[4] R.G. Gallager, Information Theory and Reliable Communication, John Wiley & Sons, New York, 1968.
[5] M.G. Kendall, A. Stuart, The Advanced Theory of Statistics, Vol. 2: Inference and Relationship, London, 1961.
[6] G. Morvai, S. Yakowitz, L. Gyorfi, Nonparametric inference for ergodic, stationary time series, Annals of Statistics 24 (1) (1996) 370–379.
[7] J. Rissanen, MDL denoising, IEEE Transactions on Information Theory 46 (7) (2000) 2537–2543.
[8] T. Roos, P. Myllymaki, J.J. Rissanen, MDL denoising revisited, IEEE Transactions on Signal Processing 57 (9) (2009) 3347–3360.
[9] D. Ryabko, Testing composite hypotheses about discrete ergodic processes, TEST 21 (2) (2012) 317–329.
[10] B. Ryabko, J. Astola, Universal codes as a basis for time series testing, Statistical Methodology 3 (2006) 375–397.
[11] B. Ryabko, D. Ryabko, Confidence sets in time—series filtering, in: Proceedings of 2011 IEEE International Symposium on Information Theory, ISIT'11, July 31–August 5, 2011, Saint-Petersburg, Russia.
[12] N. Usotskaya, B. Ryabko, Applications of information-theoretic tests for analysis of DNA sequences based on Markov chain models, Computational Statistics and Data Analysis 53 (5) (2009) 1861–1872.
[13] P.M.B. Vitányi, F.J. Balbach, R.L. Cilibrasi, M. Li, Normalized information distance, in: Information Theory and Statistical Learning, 2009, pp. 45–82.