Proceedings of the Seventh Young Statisticians Meeting
Andrej Mrvar (Editor)
Metodološki zvezki, 21, Ljubljana: FDV, 2003

Analysis of the Maximum-Likelihood Estimation of Hidden Markov Models

Gabor Molnár-Sáska[1]

Abstract

The estimation of Hidden Markov Models has attracted a lot of attention recently, see results of Legland and Mevel (2000) and Leroux (1992). The purpose of this paper is to give a view for the analysis of the maximum-likelihood estimation of HMM-s. General consistency results are compared to the new approach. The new approach is potentially useful for deriving strong approximation results, which are in turn applicable to analyze adaptive predictors.

1 Introduction

Hidden Markov Models have become a basic tool for modeling stochastic systems with a wide range of applicability in such diverse areas as robotics, telecommunication, econometrics and protein research. For a general introduction see van Schuppen.

The estimation of the dynamic of a Hidden Markov Model is a basic problem in applications. A key element in statistical analysis of HMM-s is a strong law of large numbers for the log-likelihood function. In previous works stability theory of Markov chains and the subadditive ergodic theorem were used, see Baum and Petrie (1966), Douc and Matias (2001), Legland and Mevel (2000) and Leroux (1992). Although these tools are very powerful, they do not yield a LLN with guaranteed rate of convergence. An alternative tool that has been widely used in linear system identification is theory of L-mixing processes, see Gerencsér (1989). The advantage of this approach is that, potentially a more precise characterization of the estimation error-process can be obtained, which, in turn, is crucial for the analysis of the performance of adaptive prediction.

2 Hidden Markov Models

We consider Hidden Markov Models with a general state space $\mathcal{X}$ and an observation or read-out space $\mathcal{Y}$. Both are assumed to be Polish spaces, i. e. they are complete,

[1] MTA SZTAKI, Computer and Automation Institute, Hungarian Academy of Sciences, 13-17 Kende u., Budapest 1111, Hungary.

separable metric spaces.

Definition 2.1 *The pair (X_n, Y_n) is a Hidden Markov process if (X_n) is a homogenous Markov chain, with state space $\mathcal{X}$ and the observations (Y_n) are conditionally independent given (X_n) and identically distributed.*

If $\mathcal{X}$ and $\mathcal{Y}$ are discrete then we have

$$P(Y_n = y_n, \dots Y_0 = y_o | X_n = x_n, \dots X_0 = x_0) = \prod_{i=0}^{n} P(Y_i = y_i | X_i = x_i).$$

In this case we will use the following notations

$$P(Y_k = y_k | X_k = x_k) = P(y_k | x_k),$$

$$P(Y = y | X = x) = b^x(y) \quad B(y) = \mathrm{diag}(b^x(y)).$$

It is well-known, see Gerencsér et al. (2002), that if (X_n, Y_n) is a Hidden Markov process, then $Z_n = (X_n, Y_n)$ is a Markov process.

For further notation let $Q > 0$ be the transition matrix of the unobserved Markov process (X_n), where $Q_{ij} = P(X_{n+1} = j | X_n = i)$, and let the predictive filter be defined by

$$p_{n+1}^j = P(X_{n+1} = j | Y_n, \dots Y_0).$$

Also write $p_{n+1} = (p_{n+1}^1, \dots, p_{n+1}^N)^T$. The filter process satisfies the Baum-equation,

$$p_{n+1} = \pi(Q^* B(Y_n) p_n). \tag{2.1}$$

where π is the normalizing operator $\pi(x)^i = \frac{x^i}{\sum_j x^j} \qquad x^j \geq 0, x \neq 0$ to make p_{n+1} a probability vector and * denotes the transpose operator. Here $p_0^j = P(X_0 = j)$. In practice we take an arbitrary probability vector q as initial condition. Then the solution of the Baum-equation will be denoted by $p_n(q)$. A key property of the Baum-equation is its exponential stability with respect to the initial condition. This is the content of the following theorem.

Theorem 2.1 *(Legland, Mevel) Let $Q > 0$ and let q and q' are different initial points. Then*

$$\|p_n(q) - p_n(q')\|_{TV} \leq C(1-\delta)^n,$$

where C is a finite constant, $0 < \delta < 1$ and $\| \quad \|_{TV}$ denotes the norm in total variation.

This basic property of the prediction filter is abstracted and used to derive general mixing properties of the extended process (X_n, Y_n, p_n), see Legland and Mevel (2000) and Gerencsér et al. (2002).

3 Estimation of Hidden Markov Models

This section gives a brief outline of the maximum likelihood estimation of Hidden Markov Models. Let the state space $\mathcal{X}$ be finite and let the transition matrix Q be positive. Furthermore let the read-outs be continuous, i.e. $Y_n \in \mathcal{Y}$, where $\mathcal{Y}$ is a Euclidean space. Let the Doeblin condition be satisfied for X_n and thus for the pair (X_n, Y_n). With the notation $p_n^i = P(X_n = i | Y_{n-1}, \ldots, Y_0)$ we have

$$p_{n+1} = \pi(Q^T B(Y_n) p_n) = f(Y_n, p_n).$$

Denote $f(y|j)$ the density function of Y_n given $X_n = j$. Assume

$$E|\log f(y|j)| < \infty \tag{3.1}$$

for all j. Y_ns are conditionally independent, so the common density function of $(Y_1, \ldots, Y_n)$ exists. Denote it by $p(y_n, \ldots, y_1, \Theta)$, where Θ is the parameter of the system.

Theorem 3.1 *The limit*

$$\lim_{n\to\infty} \frac{1}{n} E(-\log p(Y_1, \ldots, Y_n))$$

exists.

If the read-out space $\mathcal{Y}$ is finite and $\Theta = \Theta^*$ is the real parameter then the consistency of the ML is implied from the following general theorem.

Theorem 3.2 *(Shannon-McMillan-Breiman) Let Y_n be (strictly) stacionary, ergodic process. Then*

$$\lim_{n\to\infty} -\frac{1}{n} \log p(Y_1, \ldots, Y_n) = \lim_{n\to\infty} E(-\log p(Y_1, \ldots, Y_n))$$

exists with probability 1.

In case of the structure of Hidden Markov

Theorem 3.3 *If $E|\log f(y|j)| < \infty$ for all j, then*

$$\lim_{n\to\infty} \frac{1}{n}(-\log p(Y_1, \ldots, Y_n)) = \lim_{n\to\infty} E(-\log p(Y_1, \ldots, Y_n))$$

exists with probability 1.

In our approach we replace compactness with the Doeblin condition and use the concept of L-mixing. We achieve similar results as above.

For the likelihood estimation we need

$$\log p(y_{n-1}, \ldots y_0, \theta) = \sum_{k=1}^{n-1} \log p(y_k | y_{k-1}, \ldots y_0, \theta) + \log p(y_0, \theta),$$

and

$$\log P(y_k|y_{k-1},\dots y_0,\theta) = \sum_x \log b^x(y_k)P(x|y_{k-1},\dots,y_0,\theta) =$$

$$\sum_x \log b^x(y_k)p_k^x.$$

Write

$$g(y,p) = \sum_x \log b^x(y)p^x. \tag{3.2}$$

Thus we get

$$\log p(y_n,\dots,y_0,\Theta) = \frac{1}{N}\sum_{k=1}^{N} g(y_k,p_k) \tag{3.3}$$

In Gerencsér et al. (2002) the L-mixing property of the process $g(Y_n,p_n)$ is proved under similar technical conditions as in Legland Mevel (2000), where the geometric ergodicity is proved for the same process. Both implies the existence of the limit in (3.3).

Now L-mixing processes play a prominent role in modern theory of linear stochastic systems, and thus the latter result is directly applicable to derive a simple proof of the result of Baum and Petrie (1966). But it also provides the basic technical conditions, under which a very detailed characterization of the estimator process can be given in analogy with Gerencsér (1990).

In particular we have that for finite state-finite read-out HMM-s, parametrized by θ, the ML estimate of the true parameter θ^*, denoted by $\hat{\theta}_N$ satisfies, under simple technical conditions,

$$\hat{\theta}_N - \theta^* =$$

$$(R^*)^{-1}\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\theta}\log p(Y_n|Y_{n-1},\dots,Y_0,\theta^*) + r_n,$$

where $r_n = O_M(N^{-1})$ and R^* is the Fisher-information matrix.

A key point here is that the error term is $O_M(N^{-1})$. This ensures that all basic limit theorems, that are known for the dominant term, which is a martingale, are also valid for $\hat{\theta}_N - \theta^*$.

The finer characterization of the estimator process is not of purely academic interest: it plays a key role in adaptive prediction and model selection, see e.g., Gerencsér (1994).

References

[1] Baum, L.E. and Petrie, T. (1966): Statistical inference for probabilistic functions of finite state Markov chains, *Ann. Math. Stat.*, **37**, 1559-1563.

[2] Douc, R. and Matias, C. (2001): Asymptotics of the Maximum likelihood estimator for general Hidden Markov Models, *Bernoulli*, **7**, 381-420.

[3] Gerencsér, L. (1989): On a Class of Mixing Processes, *Stochastics*, **26**, 165-191.

[4] Gerencsér, L. (1990): On the martingale approximation of the estimation error of ARMA parameters, *Systems & Control Letters*, **15**, 417-423.

[5] Gerencsér, L. (1994): On Rissanen's predictive stochastic complexity for stationary ARMA processes, *Statistical Planning and Inference* , **41**.

[6] Gerencsér, L., Molnár-Sáska, G., Michaletzky, Gy., and Tusnády, G. (2002): New methods for the statistical analysis of Hidden Markov Models, *Proc. of the 41st. CDC.*

[7] LeGland, F. and Mevel, L. (2000): Exponential forgetting and geometric ergodicity in hidden markov models, *Mathematics of Control, Signals and Systems*, **13**, 63-93.

[8] Leroux, B.G. (1992): Maximum-likelihood estimation for Hidden Markov-models, *Stochastic Processes and their Applications*, **40**, 127-143.

[9] van Schuppen, J.H.: *Lecture Notes on Stochastic Systems.* Manuscript.