ESANN'2007 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2007, d-side publi., ISBN 2-930307-07-2.

Estimating the Number of Components in a Mixture of Multilayer Perceptrons

M. Olteanu and J. Rynkiewicz

SAMOS-MATISSE-CES Université Paris 1, UMR 8174
90 Rue de Tolbiac, 75013 Paris, France

Abstract. In this paper we are interested in estimating the number of components in a mixture of multilayer perceptrons. The penalized marginal-likelihood criterion for mixture models and hidden Markov models introduced by Keribin (2000) and, respectively, Gassiat (2002) is extended to mixtures of multilayer perceptrons for which a penalized-likelihood criterion is proposed. We prove the consistency of the BIC criterion under some hypothesis which involve essentially the bracketing entropy of the generalized score-functions class.

1 Introduction

Although linear models have been the standard tool for time series analysis for a long time, their limitations have been underlined during the past twenty years. Real data often exhibit characteristics that are not taken into account by linear models. Financial series, for instance, alternate strong and weak volatility periods, while economic series are often related to the business cycle and switch from recession to normal periods. Several solutions such as heteroscedatic ARCH, GARCH models, threshold models, multilayer perceptrons or autoregressive switching Markov models were proposed to overcome these problems.

In this paper, we consider models which allow the series to switch between regimes and more particularly we study the case of mixtures of multilayer perceptrons. In this frame, rather than using a single global model, we estimate several local models form the data. For the moment, we assume that switches between different models occur independently, the next step of this approach being to also learn how to split the input space and to consider the more general case of *gated experts* or *mixtures of experts* models (Jacobs et al., 1991). The problem we address here is how to select the number of components in a mixture of multilayer perceptrons. This is typically a problem of non-identifiability which leads to a degenerate Fisher information matrix and the classical chi-square theory on the convergence of the likelihood ratio fails to apply. One possible method to answer this problem is to consider penalized criteria. The consistence of the BIC criterion was recently proven for non-identifiable models such as mixtures of densities or hidden Markov models (Keribin, 2000 and Gassiat, 2002). We extend these results to mixtures of nonlinear autoregressive models and prove the consistency of a penalized estimate for the number of components under some good regularity conditions.

The rest of the paper is organized as follows : in Section 2 we give the definition of the general model and state sufficient conditions for regularity.

Afterwards, we introduce the penalized likelihood estimate for the number of components and state the result of consistency. Section 3 is concerned with applying the main result to mixtures of multilayer perceptrons. Some open questions, as well as some possible extensions are discussed in the conclusion.

2 Penalized likelihood estimate for the number of components in a mixture of nonlinear autoregressive models

The model - definition and regularity conditions

Throughout the paper, we shall consider that the number of lags is known and, for ease of writing, we shall set the number of lags equal to one, the extension to l time-lags being immediate. Let us consider the real-valued time series Y_t which verifies the true model

$$(1) \qquad Y_t = F^0_{X_t}\left(Y_{t-1}\right) + \varepsilon_{X_t}(t), \text{ where}$$

- X_t is a sequence of i.i.d. variables with values in a finite space $\{1,...,p_0\}$ and probability distribution π^0

- for every $i \in \{1,...,p_0\}$, $F^0_i(y)$ is a parametric nonlinear function depending on θ^0_i. We suppose throughout the rest of the paper that F^0_i are sublinear, that is they are continuous and there exist $\left(a^0_i, b^0_i\right)$ positive real numbers such that $\left|F^0_i(y)\right| \leq a^0_i |y| + b^0_i$, $y \in \mathbb{R}$

- for every $i \in \{1,...,p_0\}$, $\varepsilon_i(t)$ is an i.i.d. centered Gaussian noise with standard deviation σ^0_i.

We need some regularity conditions in order to prove the main result. Let us introduce the hypothesis

$$\textbf{(HS)} \qquad \sum_{i=1}^{p_0} \pi^0_i \left|a^0_i\right|^s < 1$$

Yao and Attali (2000) proved that under the hypothesis (HS), model (1) has a unique strictly-stationary solution Y_t, geometrically-ergodic. Let us remark that hypothesis **(HS)** does not request every component to be stationary and that it allows non-stationary "regimes" as long as they do not apper too often.

Construction of the penalized likelihood criterion

Let $\{y_1,...,y_n\}$ be an observed sample of the time series (Y_k). Then, the conditional density of y_k with respect to y_{k-1} is

$$f\left(y_k \mid y_{k-1}\right) = \sum_{i=1}^{p_0} \pi^0_i f^0_i\left(y_k - F^0_i\left(y_{k-1}\right)\right)$$

Now let us consider all possible conditional densities up to a maximal number of components $P > 0$, $\mathcal{G}_P = \bigcup_{p=1}^{P} \mathcal{G}_p$,

$$\mathcal{G}_p = \left\{g \mid g\left(y_1, y_2\right) = \sum_{i=1}^{p} \pi_i f_i\left(y_2 - F_i\left(y_1\right)\right), \pi_i \geq 0, \sum_{i=1}^{p} \pi_i = 1\right\}$$

where, for all $i = 1, ..., p$, F_i is a parametric function depending on θ_i, sublinear, and f_i is a centered Gaussian density with standard error σ_i. Throughout the following, we shall make a natural assumption on the compacteness of the parameters : **(HC)** $\{(\pi_i, \theta_i, \sigma_i), i = 1, ..., p\}$ belong to a compact set.

For every $g \in \mathcal{G}_P$ we define the number of components as

$$p(g) = min\{p \in \{1, ..., P\}, g \in \mathcal{G}_p\}$$

and let $p_0 = p(f)$ be the true number of regimes. We can now define the estimate $\hat{p}$ as the argument $p \in \{1, ..., P\}$ maximizing the penalized criterion

$$(2) \qquad T_n(p) = sup_{g \in \mathcal{G}_p} l_n(g) - a_n(p)$$

where $l_n(g) = \sum_{k=2}^{n} log\ g(y_{k-1}, y_k)$ is the log-likelihood and $a_n(p)$ is a penalty term.

Convergence of the penalized likelihood estimate

The next result is an extension of Gassiat (2002) for hidden Markov models and since the main steps of the proof are the same, we shall omit it.

Theorem 1 *: Consider the model (Y_k, X_k) defined by (1) and the penalized-likelihood criterion introduced in (2). Let us introduce the next assumptions :*

(A1) *$a_n(\cdot)$ is an increasing function of p, $a_n(p_1) - a_n(p_2) \to \infty$ when $n \to \infty$ for every $p_1 > p_2$ and $\frac{a_n(p)}{n} \to 0$ when $n \to \infty$ for every p*

(A2) *the model (Y_k, X_k) verifies the weak identifiability assumption **(HI)***

$$\sum_{i=1}^{p} \pi_i f_i(y_2 - F_i(y_1)) = \sum_{i=1}^{p_0} \pi_i^0 f_i^0\left(y_2 - F_i^0(y_1)\right) \Leftrightarrow \sum_{i=1}^{p} \pi_i \delta_{\theta_i} = \sum_{i=1}^{p_0} \pi_i^0 \delta_{\theta_i^0}$$

(A3) *the parameterization $\theta_i \to f_i(y_2 - F_i(y_1))$ is continuous for every (y_1, y_2) and there exists $m(y_1, y_2)$ an integrable map with respect to the stationary measure of (Y_k, Y_{k-1}) such that $|log(g)| < m$*

(A4) *Y_k is strictly stationary and geometrically β-mixing, and the family of generalized score functions associated to $\mathcal{G}_P$*

$$\mathcal{S} = \left\{s_g,\ s_g(y_1, y_2) = \frac{\frac{g(y_1, y_2)}{f(y_1, y_2)} - 1}{\left\|\frac{g}{f} - 1\right\|_{L^2(\mu)}},\ g \in \mathcal{G}_P,\ \left\|\frac{g}{f} - 1\right\|_{L^2(\mu)} \neq 0\right\} \subset \mathcal{L}_2(\mu)$$

where μ is the stationary measure of (Y_k, Y_{k-1}) and for every $\varepsilon > 0$

$$\mathcal{H}_{[\cdot]}(\varepsilon, \mathcal{S}, \|\cdot\|_2) = \mathcal{O}(|log\,\varepsilon|),$$

$\mathcal{H}_{[\cdot]}(\varepsilon, \mathcal{S}, \|\cdot\|_2)$ being the bracketing entropy of $\mathcal{S}$ with respect to the L_2-norm. Then, under hypothesis (A1)-(A4), (HS) et (HC), $\hat{p} \to p_0$ in probability.

3 Mixtures of multilayer perceptrons

In this section, we consider the model defined in (1) such that, for every $i \in \{1, ..., p_0\}$, F_i^0 is a multilayer perceptron. Since non-identifiability problems also arise in multilayer perceptrons (see, for instance, Rynkiewicz, 2006), we shall simplify the problem by considering one hidden layer and a fixed number of units on every layer, k. Then, we have that for every $i \in \{1, ..., p_0\}$

$$F_i^0(y) = \alpha_0^{0,i} + \sum_{j=1}^k \alpha_j^{0,i} \phi\left(\beta_{0,j}^{0,i} + \beta_{1,j}^{0,i} y\right)$$

where ϕ is the hyperbolic tangent and

$$\theta_i^0 = \left(\alpha_0^{0,i}, \alpha_1^{0,i}, ..., \alpha_k^{0,i}, \beta_{0,1}^{0,i}, \beta_{1,1}^{0,i}, ..., \beta_{0,k}^{0,i}, \beta_{1,k}^{0,i}\right)$$

is the true parameter.Let us check if the hypothesis of the main result of section 2 apply in the case of mixtures of multilayer perceptrons.

Hypothesis (HS) : The stationarity and ergodicity assumption (HS) is immediately verified since the output of every perceptron is bounded, by construction. Thus, every regime is stationary and the global model is also stationary.

Let us consider the class of all possible conditional densities up to a maximum number of components $P > 0$:

$$\mathcal{G}_P = \bigcup_{p=1}^P \mathcal{G}_p \ , \ \mathcal{G}_p = \{g \mid g(y_1, y_2) = \sum_{i=1}^p \pi_i f_i(y_2 - F_i(y_1))\}, \text{ where}$$

- $\sum_{i=1}^p \pi_i = 1$ and we may suppose quite naturally that for every $i \in \{1, ..., p\}$, $\pi_i \geq \eta > 0$
- for every $i \in \{1, ..., p\}$, F_i is a multilayer perceptron

$$F_i(y) = \alpha_0^i + \sum_{j=1}^k \alpha_j^i \phi\left(\beta_{0,j}^i + \beta_{1,j}^i y\right), \text{ where}$$

$\theta_i = \left(\alpha_0^i, \alpha_1^i, ..., \alpha_k^i, \beta_{0,1}^i, \beta_{1,1}^i, ..., \beta_{0,k}^i, \beta_{1,k}^i\right)$ belongs to a compact set.

Hypothesis (A1) : $a_n(\cdot)$ may be chosen, for instance, equal to the BIC penalizing term, $a_n(p) = \frac{1}{2} p \, log(n)$.

Hypothesis (A2)-(A3) : Since the noise is normally distributed, the weak identifiability hypothesis is verified according to the result of Teicher (1963), while assumption (A3) is a regularity condition verified by Gaussian densities.

Hypothesis (A4) : We consider the class of generalized score functions

$$\mathcal{S} = \left\{ s_g, \ s_g = \frac{\frac{g}{f} - 1}{\left\|\frac{g}{f} - 1\right\|_{L^2(\mu)}}, \ g \in \mathcal{G}_P, \ \left\|\frac{g}{f} - 1\right\|_{L^2(\mu)} \neq 0 \right\}$$

The difficult part will be to show that $\mathcal{H}_{[\cdot]}(\varepsilon, \mathcal{S}, \|\cdot\|_2) = \mathcal{O}(|log \, \varepsilon|)$ for all $\varepsilon > 0$ which, since we are on a functional space, is equivalent to prove that "the dimension" of $\mathcal{S}$ can be controlled. For $g \in \mathcal{G}_p$, let us denote $\theta = (\theta_1, ..., \theta_p)$ and $\pi = (\pi_1, ..., \pi_p)$, so that the global parameter will be $\Phi = (\theta, \pi)$ and the associated generalized score function $s_\Phi := s_g$.

Proving that a parametric family like $\mathcal{S}$ verifies the condition on the bracketing entropy is usually immediate under good regularity conditions (see, for instance, Van der Vaart (2000)). In this particular case, the problems arise when $g \to f$ and the limits in $L^2(\mu)$ of s_g have to be computed. Let us split $\mathcal{S}$ into two classes of functions. We shall consider $\mathcal{F}_0 \subset \mathcal{G}_P$ a neighbourhood of f such that it exists $\delta_\varepsilon > 0$ verifying $\mathcal{F}_0 = \left\{ g \in G_p, \left\| \frac{g}{f} - 1 \right\|_{L^2(\mu)} \leq \delta_\varepsilon \right\}$ and $\mathcal{S}_0 = \{s_g,\ g \in \mathcal{F}_0\}$. On $\mathcal{S} \setminus \mathcal{S}_0$, it can be easily seen that

$$\left\| \frac{\frac{g_1}{f} - 1}{\left\| \frac{g_1}{f} - 1 \right\|_{L^2(\mu)}} - \frac{\frac{g_2}{f} - 1}{\left\| \frac{g_2}{f} - 1 \right\|_{L^2(\mu)}} \right\|_{L^2(\mu)} \leq \frac{2}{\delta} \left\| \frac{g_1}{f} - \frac{g_2}{f} \right\|_{L^2(\mu)}$$

and we get that $\mathcal{N}_{[]}(\varepsilon, \mathcal{S} \setminus \mathcal{S}_0, \|\cdot\|_2) = \mathcal{O}\left(\frac{1}{\delta\varepsilon}\right)^{3(k+1)P}$, where $\mathcal{N}_{[]}(\varepsilon, \mathcal{S} \setminus \mathcal{S}_0, \|\cdot\|_2)$ is the number of ε-brackets necessary to cover $\mathcal{S} \setminus \mathcal{S}_0$ and the bracketing entropy is computed as $log\mathcal{N}_{[]}(\varepsilon, \mathcal{S} \setminus \mathcal{S}_0, \|\cdot\|_2)$.

As for $\mathcal{S}_0$, the idea is to reparameterize the model in a convenient manner which will allow a Taylor expansion around the identifiable part of the true value. For that, we shall use a slight modification of the method proposed by Liu and Shao (2003). Let us remark that when $\frac{g}{f} - 1 = 0$, the weak identifiability hypothesis **(A2)** and the fact that for every $i \in \{1, ..., p\}$, $\pi_i \geq \eta > 0$, imply that there exists a vector $t = (t_i)_{0 \leq i \leq p_0}$ such that $0 = t_0 < t_1 < ... < t_{p_0} = p$ and, modulo a permutation, Φ can be rewritten as follows : $\theta_{t_{i-1}+1} = ... = \theta_{t_i} = \theta_i^0$, $\sum_{j=t_{i-1}+1}^{t_i} \pi_j = \pi_i^0$, $i \in \{1, ..., p_0\}$. With this remark, one can define in the general case $s = (s_i)_{1 \leq i \leq p_0}$ and $q = (q_j)_{1 \leq j \leq p}$ so that, for every $i \in \{1, ..., p_0\}$, $j \in \{t_{i-1} + 1, ..., t_i\}$,

$$s_i = \sum_{j=t_{i-1}+1}^{t_i} \pi_j - \pi_i^0,\ q_j = \frac{\pi_j}{\sum_{l=t_{i-1}+1}^{t_i} \pi_l}$$

and the new parameterization will be $\Theta_t = (\phi_t, \psi_t)$,

$\phi_t = \left((\theta_j)_{1 \leq j \leq p}, (s_i)_{1 \leq i \leq p_0 - 1} \right)$, $\psi_t = (q_j)_{1 \leq j \leq p}$, with ϕ_t containing all the identifiable parameters of the model and ψ_t the non-identifiable ones. Then, for $g = f$, we will have that

$$\phi_t^0 = (\underbrace{\theta_1^0, ..., \theta_1^0}_{t_1}, ..., \underbrace{\theta_{p_0}^0, ..., \theta_{p_0}^0}_{t_{p_0} - t_{p_0 - 1}}, \underbrace{0, ..., 0}_{p_0 - 1})^T$$

This reparameterization allows to write a second-order Taylor expansion of $\frac{g}{f} - 1$ at ϕ_t^0. For ease of writing, we shall first denote

$$g_j(y_1, y_2) = g_{\theta_j}(y_1, y_2) = \frac{f_j(y_2 - F_j(y_1))}{\sum_{i=1}^{p_0} \pi_i^0 f_i^0 \left(y_2 - F_i^0(y_1)\right)} - 1$$

Then, the density ratio becomes :

$$\frac{g}{f} - 1 = \sum_{i=1}^{p_0 - 1} \left(s_i + \pi_i^0\right) \sum_{j=t_{i-1}+1}^{t_i} q_j g_j + \left(\pi_{p_0}^0 - \sum_{i=1}^{p_0 - 1} s_i\right) \sum_{j=t_{p_0 - 1}+1}^{t_{p_0}} q_j g_j$$

By remarking that when $\phi_t = \phi_t^0$, $\frac{g}{f}$ does not vary with ψ_t, we will study the variation of this ratio in a neighbourhood of ϕ_t^0 and for fixed ψ_t. Assuming that $\left(g_j\right)_{1\leq j\leq p}$, $\left(g'_j\right)_{1\leq j\leq p}$, $\left(g''_j\right)_{1\leq j\leq p}$ and $\left(g'''_j\right)_{1\leq j\leq p}$, where

$$g'_j := \frac{\partial g_j}{\partial \theta_j}\left(\phi_t^0,\psi_t\right),\ g''_j := \frac{\partial^2 g_j}{\partial \theta_j^2}\left(\phi_t^0,\psi_t\right),\ g'''_j := \frac{\partial^3 g_j}{\partial \theta_j^3}\left(\phi_t^0,\psi_t\right)$$

are linearly independent in $L^2\left(\mu\right)$, one can prove the following :

Proposition 4 *: Let us denote* $D\left(\phi_t,\psi_t\right) = \left\|\frac{g(\phi_t,\psi_t)}{f} - 1\right\|_{L^2(\mu)}$. *For any fixed* ψ_t*, there exists the second-order Taylor expansion at* ϕ_t^0 *:*

$$\frac{g}{f} - 1 = \left(\phi_t - \phi_t^0\right)^T g'_{\left(\phi_t^0,\psi_t\right)} + \frac{1}{2}\left(\phi_t - \phi_t^0\right)^T g''_{\left(\phi_t^0,\psi_t\right)}\left(\phi_t - \phi_t^0\right) + o\left(D\left(\phi_t,\psi_t\right)\right),$$

$$\left(\phi_t - \phi_t^0\right)^T g'_{\left(\phi_t^0,\psi_t\right)} + \frac{1}{2}\left(\phi_t - \phi_t^0\right)^T g''_{\left(\phi_t^0,\psi_t\right)}\left(\phi_t - \phi_t^0\right) = 0 \Leftrightarrow \phi_t = \phi_t^0$$

Using the Taylor expansion above, one can show that $\mathcal{N}_{[]}\left(\varepsilon, \mathcal{S}_0, \|\cdot\|_2\right) = \mathcal{O}\left(\frac{1}{\varepsilon}\right)^{Ckp_0}$ and the assumptions of Theorem 1 are verified.

4 Conclusion and future work

We have proven the consistency of the BIC criterion for estimating the number of components in a mixture of multilayer perceptrons. In our opinion, two important directions are to be studied in the future. The case of mixtures should be extended to the general case of gated experts which allow the probability distribution of the multilayer perceptrons to depend on the input and thus, to learn how to split the input space. The second possible extension should remove the hypothesis of a fixed number of units on the hidden layer. The problem of estimating the number of hidden units in one multilayer perceptron was solved in Rynkiewicz (2006), but it would be interesting to mix the two results and prove the consistency of a penalized criterion when there is a double non-identifiability problem : number of experts and number of hidden units.

References

[1] Gassiat E. (2002) Likelihood ratio inequalities with applications to various mixtures, *Ann. Inst. Henri Poincaré*, 38, 897-906

[2] Jacobs R.A., Jordan M.I., Nowlan S.J., Hinton G.E. (1991) Adaptive mixtures of local experts, *Neural Computation*, 3, 79-87

[3] Keribin C. (2000) Consistent estimation of the order of mixture models, *Sankhya : The Indian Journal of Statistics*, 62, 49-66

[4] Liu X., Shao Y. (2003) Asymptotics for likelihood ratio tests under loss of identifiability, *The Annals of Statistics*, 31(3), 807-832

[5] Rynkiewicz J. (2006) Consistent estimation of the architecture of multilayer perceptrons, *ESANN'2006 Proceedings*, d-side publi., 149-154

[6] Teicher H. (1963) Identifiability of finite mixtures, *Ann. Math. Statist.*, 34(2), 1265-1269

[7] Van der Vaart A.W. (2000) *Asymptotic Statistics*, Cambridge University Press

[8] Yao J.F., Attali J.G. (2000) On stability of nonlinear AR processes with Markov switching, *Advances in Applied Probability*, 32 (2), 394-407