ESANN 2010 proceedings, European Symposium on Artificial Neural Networks - Computational Intelligence and Machine Learning. Bruges (Belgium), 28-30 April 2010, d-side publi., ISBN 2-930307-10-2.

Asymptotic properties of mixture-of-experts models

M. Olteanu and J. Rynkiewicz

Universite Paris 1 - SAMOS-CES
90 Rue de Tolbiac, 75013 Paris - France

Abstract. The statistical properties of the likelihood ratio test statistic (LRTS) for mixture-of-expert models are addressed in this paper. This question is essential when estimating the number of experts in the model. Our purpose is to extend the existing results for mixtures (Liu and Shao, 2003) and mixtures of multilayer perceptrons (Olteanu and Rynkiewicz, 2008). In this paper we study a simple example which embodies all the difficulties arising in such models. We find that in some cases the LRTS diverges but, with additional assumptions, the behavior of such models can be totally explicated.

1 Introduction

Derived from neural networks literature, Mixtures of Experts (ME) (Jacobs et. al., 1991) and Hierarchical Mixtures of Experts (HME) (Jordan and Jacobs, 1994) generalize linear regression models. HME are mixtures of "experts" (for example, linear regression models) organized in a tree-structured network. The network assigns a weight which, unlike mixture regression models, may depend on the input x to each expert and then produces an output which combines the outputs produced by all experts according to their weights. The ME discussed in this paper is a particular case of HME, where the network has only one layer.

The conditional density of a ME can be generally written as:

$$g\left(y|x,\phi\right) = \sum_{i=1}^{p} \pi_{\nu_i}(x) g_{\theta_i}(y|x),$$

where $\phi = \left(\nu_1^T, ..., \nu_p^T, \theta_1^T, ..., \theta_p^T\right)$ is the parameter of the model. Usually, the weights or "gating functions" are chosen to be logistic type

$$\pi_{\nu_i}(x) = \frac{\exp\left(\nu_i^T x\right)}{\sum_{j=1}^{p} \exp\left(\nu_j^T x\right)},$$

while g_θ may be Poisson, Binomial or Gaussian distributions.

When the model is assumed to be correctly specified, the maximum likelihood estimates converge to the true values of the parameters and are normally distributed (Jiang and Tanner, 1999). However, the true model is not usually known and the true parameter is unidentifiable. This paper studies the asymptotic behavior of the likelihood ratio test statistic (LRTS) for mixtures of experts and extends the results for mixtures of Liu and Shao (2003). In Section 2, we present the model and a simple example. An example of divergence is given in Section 3, while Section 4 describes the convergence of the LRTS under some additional assumptions.

2 The model and a simple example

Let $(X_k, Y_k)_{k\in\mathbb{Z}}$ be a sequence of independent and identically distributed random vectors defined on a probability space $(\Omega, \mathcal{K}, \mathbb{P})$. Let $\mathcal{P} = \{g_\theta, \theta \in \Theta\}$ be a set of densities with respect to some positive measure λ, where Θ is a finite-dimensional set. Let us consider an observed sample $\{(x_1, y_1), ..., (x_n, y_n)\}$ of the sequence (X_k, Y_k). For every x_k, the true density of Y_k conditionally to $X_k = x_k$ is

$$g^0\left(y_k \mid x_k\right) = \textstyle\sum_{i=1}^{p_0} \pi_{\nu_i{}^0}(x) g_{\theta_i{}^0}\left(y_k \mid x_k\right),$$

where $g_{\theta^0} \in \mathcal{P}$, $\pi_{\nu_i{}^0}(x) \geq 0$, $\sum_{i=1}^{p_0} \pi_{\nu_i{}^0}(x) = 1$ and $\phi_0 = \left(\theta_1^0, ..., \theta_{p_0}^0, \nu_1^0, ..., \nu_{p_0}^0\right)^T$ is the global parameter of the model. Let us remark that this model is the general parameterization of mixtures of experts. In the next section, a simple example of such model is studied.

2.1 Simple mixture of experts

Let $\mathcal{G}$ be the set of possible conditional densities:

$$\mathcal{G} = \{g\left(y \mid x\right) = \pi_\nu(x) g_b\left(y \mid x\right) + (1 - \pi_\nu(x)) f\left(y \mid x\right), \pi_\nu(x) \in [0;1], g_b \in \mathcal{P}\}$$

with $\mathcal{P} = \left\{g_b(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-bx)^2}, b \in \Theta \subset \mathbb{R}\right\}$ the set of conditional densities and $f\left(y \mid x\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2}$. This model is clearly a particular case of the general mixture of expert model and is a simple example of mixture of regressions with Gaussian noise. Let

$$l_n\left(g\right) = \textstyle\sum_{k=2}^{n} \ln\, g\left(y_k \mid x_k\right)$$

be the conditional log-likelihood function of $((x_1, y_1), \cdots, (x_n, y_n))$. We want to know whether the true model is really a mixture regression model (i.e. $b \neq 0$ and $\pi_\nu(x) \neq 0$) or the observations are independent ($b = 0$ or $\forall x, \pi_\nu(x) = 0$). We need to look at the likehood ratio test statistic (LRTS) to answer this question. The LRTS is defined as:

$$2\lambda_n = 2\left(\sup_{g\in\mathcal{G}} \ln(g) - \ln(f)\right) = 2 \sup_{g\in\mathcal{G}} \frac{\pi_\nu(x) g_b\left(y \mid x\right) + (1 - \pi_\nu(x)) f\left(y \mid x\right)}{f\left(y \mid x\right)} \quad (1)$$

For regular statistical models, the LRTS converges to a χ^2 distribution. This is no longer the case with this model. Let us first recall a result which gives an approximation of the LRTS.

2.2 Approximation of the LRTS

First, we have to introduce some definitions and properties.

- The extended set of score-functions $\mathcal{S}$ is defined as:

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

- Consider the extended set of score-functions $\mathcal{S}$ endowed with the norm $\|\cdot\|_{L^2(\mu)}$. For every $\varepsilon > 0$, we define an ε-bracket by $[l,\, u] = \{f \in \mathcal{F},\, l \leq f \leq u\}$ such that $\|u - l\|_{L^2(\mu)} < \varepsilon$. The ε-bracketing entropy is
$$\mathcal{H}_{[\cdot]}\left(\varepsilon, \mathcal{S}, \|\cdot\|_{L^2(\mu)}\right) = \ln\left(\mathcal{N}_{[\cdot]}\left(\varepsilon, \mathcal{S}, \|\cdot\|_{L^2(\mu)}\right)\right),$$
where $\mathcal{N}_{[\cdot]}\left(\varepsilon, \mathcal{S}, \|\cdot\|_{L^2(\mu)}\right)$ is the minimum number of ε-brackets necessary to cover $\mathcal{S}$.
- With the previous notations, we introduce the following assumption **(B)**: for all $\eta > 0$, denote $\mathcal{G}_\eta := \left\{g \in \mathcal{G}, \|\frac{g}{f} - 1\|_2 \leq \eta\right\}$ and $\mathcal{S}_\eta := \{s \in \mathcal{S}, g \in \mathcal{G}_\eta\}$. Assume that $\mathcal{G}$ is Glivenko-Cantelli and that there exists $\eta > 0$ such that
$$\int_0^1 \sqrt{\mathcal{H}_{[\cdot]}\left(\varepsilon, \mathcal{S}_\eta, \|\cdot\|_{L^2(\mu)}\right)} d\varepsilon < \infty.$$
Then the set $\mathcal{S}_\eta$ is Donsker under **(B)**.
- Let us also define the limit-set of scores $\mathcal{D}$
$$\left\{d \in \mathbb{L}^2(\mu) \mid \exists (g_n) \in \mathcal{G}, \|\tfrac{g_n - f}{f}\|_{\mathbb{L}^2(\mu)} \xrightarrow[n\to\infty]{} 0, \|d - s_{g_n}\|_{\mathbb{L}^2(\mu)} \xrightarrow[n\to\infty]{} 0\right\}.$$
By putting $g_t = g_n$ for $t \in [0,1]$ and $n \leq \frac{1}{t} < n+1$, we obtain that, for all $d \in \mathcal{D}$, there exists a parametric path $(g_t)_{0\leq t\leq 1}$ such that $\forall t \in [0,1]$, $g_t \in \mathcal{G}$, $t \to \|\frac{g_t - f}{f}\|_{\mathbb{L}^2(\mu)}$ is continuous in 0, $\|\frac{g_t - f}{f}\|_{\mathbb{L}^2(\mu)} \xrightarrow[t\to 0]{} 0$ and $\|d - s_{g_t}\|_{\mathbb{L}^2(\mu)} \xrightarrow[t\to 0]{} 0$.

The following theorem can be stated (Gassiat, 2002):
Theorem 1: *Under the assumption (B),*
$$2\lambda_n = \sup_{d\in\mathcal{D}} \left(\max\left\{\frac{1}{\sqrt{n}}\sum_{i=2}^n d(Y_i, X_i); 0\right\}\right)^2 + o_P(1)$$

In order to derive the behaviour of the LRTS, two cases have to be analyzed. The first one is when $\exists \delta > 0$ such that $\forall \nu$, $E\left[\pi_\nu(X)\right] \geq \delta$. The second one is if one can find a sequence of parameters $\nu_1, \cdots, \nu_n$ such that $\lim_{n\to\infty} E\left[\pi_{\nu_n}(X)\right] = 0$.

3 Divergence of LRTS

The LRTS can be divergent if there exists a sequence of parameters $\nu_1, \cdots, \nu_n, \cdots$ such that $\lim_{n\to\infty} E\left[\pi_{\nu_n}(X)\right] = 0$. Indeed, for such sequence we can have $\|\ln(g) - \ln(f)\| \to 0$ with $b \neq 0$.

For sake of simplicity, assume that the probability function $\pi_\nu(X)$ is constant. Then, if the quantity $\left\|\exp\left(-\frac{b^2}{2}X^2 + bYX\right) - 1\right\|_{L^2(\mu)}$ is finite, the score functions are well defined. Let us study

$$\begin{array}{l}
\left\| \exp\left(-\frac{b^2}{2}X^2 + bYX\right) - 1 \right\|^2_{L^2(\mu)} = \\
\frac{1}{2\pi}\int\int\left(\exp\left(-\frac{b^2}{2}x^2+byx\right)-1\right)^2 \exp\left(-\frac{1}{2}x^2\right)\exp\left(-\frac{1}{2}y^2\right)dxdy = \\
\frac{1}{2\pi}\int\int\left(\exp\left(-b^2x^2+2byx\right) - 2\exp\left(-\frac{b^2}{2}x^2+byx\right)+1\right) \\
\exp\left(-\frac{1}{2}x^2\right)\exp\left(-\frac{1}{2}y^2\right)dxdy
\end{array}$$

The integral of the dominant term (the first) is:

$$\begin{array}{rcl}
I(b) & = & \frac{1}{2\pi}\int\int\exp\left(-b^2x^2+2byx\right)\exp\left(-\frac{1}{2}x^2\right)\exp\left(-\frac{1}{2}y^2\right)dxdy \\
& = & \frac{1}{2\pi}\int\int\exp\left(-\left(b^2+\frac{1}{2}\right)x^2+2bxy-\frac{1}{2}y^2\right)dxdy \\
& = & \frac{1}{2\pi}\int\int\exp\left(-\left(\sqrt{b^2+\frac{1}{2}}x-\frac{b}{\sqrt{b^2+\frac{1}{2}}}y\right)^2-\left(\frac{1}{2}-\frac{b^2}{b^2+\frac{1}{2}}\right)y^2\right)dxdy \\
& = & \frac{\sqrt{2b^2+1}}{\sqrt{2\pi}}\int\exp\left(-\left(\frac{1}{2}-\frac{b^2}{b^2+\frac{1}{2}}\right)y^2\right)dy
\end{array}$$

Finally for $-\frac{1}{\sqrt{2}} < b < \frac{1}{\sqrt{2}}$,

$$\left\|\exp\left(-\frac{b^2}{2}x^2+byx\right)-1\right\|_{L^2(\mu)} < +\infty$$

and the score function is well defined. The set of limit score functions contains the score functions:

$$\left\{\nu_b(x,y) = \frac{\frac{g_b(y,x)}{f(y,x)}}{\|\frac{g_b(y,x)}{f(y,x)}\|_{L^2(\mu)}}, b \in]-\frac{1}{\sqrt{2}};\frac{1}{\sqrt{2}}[\right\}$$

Note that the distribution of the LRTS λ_n for a finite number of possible parameters $b_1, \cdots, b_m$ will always converge to a m-dimensional normal distribution with covariance $\left(E\left(\nu_{b_i}(x,y)\,\nu_{b_j}(x,y)\right)\right)_{1\leq i,j\leq m}$. Suppose that an arbitrary number of“almost” uncorrelated random variables in C can be found, then λ_n can take an arbitrarily large value since the maximum of m independent samples from standard normal distribution is approximately $\sqrt{2\log m}$. Hence, Fukumizu (2003) has shown that if a sequence $b_1, \cdots, b_m, \cdots$ exists so that

$$\lim_{m\to\infty}\nu_{b_m}(x,y)\xrightarrow{P}0$$

then the likelihood ratio T_n diverges to infinite. Here, we get

$$\lim_{b\to\frac{1}{\sqrt{2}}, b<\frac{1}{\sqrt{2}}}\left\|\exp\left(-\frac{b^2}{2}x^2+byx\right)-1\right\|_{L^2(\mu)} = +\infty$$

So, for each sphere B of $\mathbb{R}^2$, centered on the origin, if $(x,y)\in B$:

$$\lim_{b\to\frac{1}{\sqrt{2}}, b<\frac{1}{\sqrt{2}}}\frac{\exp\left(-\frac{b^2}{2}x^2+byx\right)-1}{\left\|\exp\left(-\frac{b^2}{2}x^2+byx\right)-1\right\|_{L^2(\mu)}} = 0$$

and $\frac{\exp\left(-\frac{b^2}{2}x^2+byx\right)-1}{\left\|\exp\left(-\frac{b^2}{2}x^2+byx\right)-1\right\|_{L^2(\mu)}}$ converges to 0 in probability for $b\to\frac{1}{\sqrt{2}}, b<\frac{1}{\sqrt{2}}$. With the choice $b_m = \frac{1}{\sqrt{2}}-\frac{1}{m}$, we get $\lim_{m\to\infty}\nu_{b_m}(x,y)\xrightarrow{P}0$ and the LRTS is divergent.

4 Convergence of LRTS

In this section we suppose that $(\exists)\delta > 0$ such that $(\forall)\nu$, $\mathbb{E}\left(\pi_\nu(X)\right) \geq \delta > 0$. Since $\pi_\nu(X) \geq 0$, then $(\exists)A \subseteq \mathbb{R}$ such that $\lambda(A) = \eta > 0$ and $\pi_\nu(x) > \delta$ for any $x \in A$. The generalized score-function can be rewritten using the following:

$$\begin{aligned} \tfrac{g}{g_0} - 1 &= \tfrac{\pi_\nu(x)g_\theta(y|x)+(1-\pi_\nu(x))f(y|x)}{f(y|x)} - 1 \\ &= \pi_\nu(x)\left(\tfrac{g_\theta(y|x)}{f(y|x)} - 1\right) \\ s_g = \tfrac{\frac{g}{g_0}-1}{\|\frac{g}{g_0}-1\|_{L^2}} &= \tfrac{\pi_\nu(x)\left(\frac{g_\theta(y|x)}{f(y|x)}-1\right)}{\|\pi_\nu(x)\left(\frac{g_\theta(y|x)}{f(y|x)}-1\right)\|_{L^2}} \end{aligned}$$

The model is parameterized by $\phi = (\theta, \nu) \in \Theta \times V \subseteq \mathbb{R}^2$ compact set and θ_0 belongs to the interior of Θ. Since $\mathbb{E}\left(\pi_\nu(X)\right) \geq \delta > 0$, we have that $g = g_0 \Leftrightarrow \theta = \theta_0$. Thus, the model is identifiable in θ and unidentifiable in ν. For any fixed $\nu \in V$, we have the following Taylor expansion around θ_0:

$$l_{(\theta,\nu)} - 1 = (\theta - \theta_0)\,\tfrac{\partial}{\partial\theta} l_{(\theta_0,\nu)} + o\left(|\theta - \theta_0|\right)$$

where $l_{(\theta,\nu)} = \frac{g_\theta}{f}$. Hence,

$$\begin{aligned} s_g = s_{\phi=(\theta,\nu)} &= \tfrac{\frac{g}{g_0}-1}{\|\frac{g}{g_0}-1\|_{L^2}} \\ &= \tfrac{\pi_\nu(x)\left[(\theta-\theta_0)\frac{\partial}{\partial\theta}l_{(\theta_0,\nu)}+o(|\theta-\theta_0|)\right]}{\|\pi_\nu(x)\left[(\theta-\theta_0)\frac{\partial}{\partial\theta}l_{(\theta_0,\nu)}+o(|\theta-\theta_0|)\right]\|_{L^2}} \\ &= \beta\tfrac{\pi_\nu(x)\left[\frac{\partial}{\partial\theta}l_{(\theta_0,\nu)}+o(1)\right]}{\|\pi_\nu(x)\left[\frac{\partial}{\partial\theta}l_{(\theta_0,\nu)}+o(1)\right]\|_{L^2}} \end{aligned}$$

where $|\beta| = 1$.

In the Gaussian case, $g_\theta(y|x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(y - \theta x\right)^2\right)$, the first derivative of $l_{(\theta,\nu)}$ is $\frac{\partial}{\partial\theta}l_{(\theta_0,\nu)}(x,y) = x\left(y - \theta_0 x\right)$, hence the directional score functions are not linearly correlated ($\beta\pi_\nu(x)\frac{\partial}{\partial\theta}l_{(\theta_0,\nu)}(x,y) \neq 0$ for any β and $\nu \in V$) and we may apply Lemma 4.1 in Liu and Shao (2003). According to this lemma, when $|\theta - \theta_0| \to 0$, the set of limit score functions is $\mathcal{F}$, where

$$\mathcal{F} = \left\{\Omega\left(\beta\pi_\nu(x)\tfrac{\partial}{\partial\theta}l_{(\theta_0,\nu)}(x,y)\right), |\beta| = 1, \nu \in V\right\}$$

and $\Omega(f) = \frac{f}{\|f\|_{L^2}}$. According to Theorem 3.1 in Liu and Shao (2003), the LRTS satisfies:

$$\lim 2\lambda_n = \sup\nolimits_{s_g \in \mathcal{F}}\left(W_{s_g} \vee 0\right)^2$$

where $\left\{W_{s_g}, s_g \in \mathcal{F}\right\}$ is a centered Gaussian process with continuous sample paths and covariance kernel $\mathbb{E}\left(W_{s_1}W_{s_2}\right) = \mathbb{E}\left(s_1 s_2\right)$. In our case,

$$\mathcal{F} = \left\{\Omega\left(\beta\pi_\nu(x)x\left(y - \theta_0 x\right)\right), |\beta| = 1, \nu \in V\right\}$$

and

$$\mathbb{E}\left(s_1 s_2\right) = \tfrac{\beta_1\beta_2\mathbb{E}\left(X^2\pi_{\nu_1}(X)\pi_{\nu_2}(X)\right)}{\|\beta_1\pi_{\nu_1}(X)X(Y-\theta_0 X)\|_{L^2}\|\beta_2\pi_{\nu_2}(X)X(y-\theta_0 X)\|_{L^2}}$$

5 Conclusion

If the number of experts is overestimated, mixture-of-expert models are no longer regular since the Fisher information matrix is singular. In this case, the LRTS provides an insight on the overfitting of the model. Although for regular models overfitting is low (some degrees of freedom of a χ^2 law), the simple example studied above shows that overfitting is more significant in the case mixture-of-expert models. The example of this paper illustrates the two main behaviors that one can expect: moderate overfitting if the mixing probabilities are bounded from below and strong overfitting if the mixing probabilities can be as small as possible. Generalization to general mixture-of-experts models should not be too difficult.

References

[1] Dacunha-Castelle D., Gassiat E. (1997a) The estimation of the order of a mixture model, *Bernoulli*, **3**, 279-299

[2] Dacunha-Castelle D., Gassiat E. (1997b) Testing in locally conic models, *ESAIM Prob. and Stat.*, **1**, 285-317

[3] Dacunha-Castelle D., Gassiat E. (1999) Testing the order of a model using locally conic parameterization: population mixtures and stationary ARMA processes, *The Annals of Statistics*, **27(4)** , 1178-1209

[4] Fukumizu K. (2003) Likelihood ratio of unidentifiable models and multilayer neural networks, *Ann. Statist.* , **31**, 833-851

[5] Gassiat E., Keribin C. (2000) The likelihood ratio test for the number of components in a mixture with Markov regime, *ESAIM P&S*, **4**, 25-52

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

[7] Henna J. (1985) On estimating the number of constituents of a finite mixture of continuous distributions, *Ann. Inst. Statist. Math.*, **37**, 235-240

[8] Izenman A.J., Sommer C. (1988) Philatelic mixtures and multivariate densities, *Journal of the American Stat. Assoc.*, **83**, 941-953

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

[10] Jiang W., Tanner M.A. (1999) On the asymptotic normality of Hierarchical Mixtures-of-Experts for Generalized Linear Models, *IEEE Trans. on Information Theory*, **46**, 1005-1013

[11] Jordan M.I., Jacobs R.A. (1994) Hierarchical mixtures of experts and the EM algorithm, *Neural Comp.*, **6**, 181-214

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

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

[14] Olteanu M., Rynkiewicz J. (2008) Estimating the number of components in a mixture of multilayer perceptrons, *Neurocomputing / EEG Neurocomputing*, **71**, 1321-1329

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