Testing the number of parameters of multidimensional MLP

Joseph Rynkiewicz

SAMOS - MATISSE
Université de Paris I
72 rue Regnault, 75013 Paris, France
(e-mail: `joseph.rynkiewicz@univ-paris1.fr`)

Abstract. This work concerns testing the number of parameters in one hidden layer multilayer perceptron (MLP). For this purpose we assume that we have identifiable models, up to a finite group of transformations on the weights, this is for example the case when the number of hidden units is know. In this framework, we show that we get a simple asymptotic distribution, if we use the logarithm of the determinant of the empirical error covariance matrix as cost function.
Keywords: Multilayer Perceptron, Statistical test, Asymptotic distribution.

1 Introduction

Consider a sequence $(Y_t, Z_t)_{t \in \mathbb{N}}$ of i.i.d.[1] random vectors (i.e. identically distributed and independents). So, each couple (Y_t, Z_t) has the same law that a generic variable $(Y, Z) \in \mathbb{R}^d \times \mathbb{R}^{d'}$.

1.1 The model

Assume that the model can be written

$$Y_t = F_{W^0}(Z_t) + \varepsilon_t$$

where

- F_{W^0} is a function represented by a one hidden layer MLP with parameters or weights W^0 and sigmoidal functions in the hidden unit.
- The noise, $(\varepsilon_t)_{t \in \mathbb{N}}$, is sequence of i.i.d. centered variables with unknown invertible covariance matrix $\Gamma(W^0)$. Write ε the generic variable with the same law that each ε_t.

Notes that a finite number of transformations of the weights leave the MLP functions invariant, these permutations form a finite group (see [Sussman, 1992]). To overcome this problem, we will consider equivalence classes of

---

[1] It is not hard to extend all what we show in this paper for stationary mixing variables and so for time series

MLP : two MLP are in the same class if the first one is the image by such transformation of the second one, the considered set of parameter is then the quotient space of parameters by the finite group of transformations.

In this space, we assume that the model is identifiable, this can be done if we consider only MLP with the true number of hidden units (see [Sussman, 1992]). Note that, if the number of hidden units is over-estimated, then such test can have very bad behavior (see [Fukumizu, 2003]). We agree that the assumption of identifiability is very restrictive, but we want emphasize the fact that, even in this framework, classical test of the number of parameters in the case of multidimensional output MLP is not satisfactory and we propose to improve it.

1.2   testing the number of parameters

Let q be an integer lesser than s, we want to test "$H_0 : W \in \Theta_q \subset \mathbb{R}^q$" against "$H_1 : W \in \Theta_s \subset \mathbb{R}^s$", where the sets Θ_q and Θ_s are compact. H_0 express the fact that W belongs to a subset of Θ_s with a parametric dimension lesser than s or, equivalently, that $s - q$ weights of the MLP in Θ_s are null. If we consider the classic cost function : $V_n(W) = \sum_{t=1}^n \|Y_t - F_W(Z_t)\|^2$ where $\|x\|$ denotes the Euclidean norm of x, we get the following statistic of test :

$$S_n = n \times \left(\min_{W \in \Theta_q} V_n(W) - \min_{W \in \Theta_s} V_n(W) \right)$$

It is shown in [Yao, 2000], that S_n converges in law to a ponderated sum of χ_1^2

$$S_n \xrightarrow{\mathcal{D}} \sum_{i=1}^{s-q} \lambda_i \chi_{i,1}^2$$

where the $\chi_{i,1}^2$ are $s - q$ i.i.d. χ_1^2 variables and λ_i are strictly positives values, different of 1 if the true covariance matrix of the noise is not the identity. So, in the general case, where the true covariance matrix of the noise is not the identity, the asymptotic distribution is not known, because the λ_i are not known and it is difficult to compute the asymptotic level of the test.

To overcome this difficulty we propose to use instead the cost function

$$U_n(W) := \ln \det \left(\frac{1}{n} \sum_{t=1}^n (Y_t - F_W(Z_t))(Y_t - F_W(Z_t))^T \right). \tag{1}$$

we will show that, under suitable assumptions, the statistic of test :

$$T_n = n \times \left(\min_{W \in \Theta_q} U_n(W) - \min_{W \in \Theta_s} U_n(W) \right) \tag{2}$$

will converge to a classical χ_{s-q}^2 so the asymptotic level of the test will be very easy to compute. The sequel of this paper is devoted to the proof of this property.

2   Asymptotic properties of T_n

In order to investigate the asymptotic properties of the test we have to prove the consistency and the asymptotic normality of $\hat{W}_n = \arg\min_{W \in \Theta_s} U_n(W)$. Assume, in the sequel, that ε has a moment of order at least 2 and note

$$\Gamma_n(W) = \frac{1}{n} \sum_{t=1}^{n} (Y_t - F_W(Z_t))(Y_t - F_W(Z_t))^T$$

remark that these matrix $\Gamma_n(W)$ and it inverse are symmetric. in the same way, we note $\Gamma(W) = \lim_{n \to \infty} \Gamma_n(W)$, which is well defined because of the moment condition on ε

2.1   Consistency of $\hat{W}_n$

First we have to identify contrast function associated to $U_n(W)$

<u>Lemma 1</u>
$$U_n(W) - U_n(W^0) \overset{a.s.}{\to} K(W, W^0)$$

with $K(W, W^0) \geq 0$ and $K(W, W^0) = 0$ if and only if $W = W^0$.

Proof :   By the strong law of large number we have

$U_n(W) - U_n(W^0) \overset{a.s.}{\to} \ln\det(\Gamma(W)) - \ln\det(\Gamma(W^0)) = \ln\frac{\det(\Gamma(W))}{\det(\Gamma(W^0))} =$
$\ln\det\left(\Gamma^{-1}(W^0)\left(\Gamma(W) - \Gamma(W^0)\right) + I_d\right)$

where I_d denotes the identity matrix of $\mathbb{R}^d$. So, the lemme is true if $\Gamma(W) - \Gamma(W^0)$ is a positive matrix, null only if $W = W^0$. But this property is true since

$\Gamma(W) = E\left((Y - F_W(Z))(Y - F_W(Z))^T\right) =$
$E\left((Y - F_{W^0}(Z) + F_{W^0}(Z) - F_W(Z))(Y - F_{W^0}(Z) + F_{W^0}(Z) - F_W(Z))^T\right) =$
$E\left((Y - F_{W^0}(Z))(Y - F_{W^0}(Z))^T\right) +$
$E\left((F_{W^0}(Z) - F_W(Z))(F_{W^0}(Z) - F_W(Z))^T\right) =$
$\Gamma(W^0) + E\left((F_{W^0}(Z) - F_W(Z))(F_{W^0}(Z) - F_W(Z))^T\right) \blacksquare$

We deduce then the theorem of consistency :

<u>Theorem 1</u> *If $E\left(\|\varepsilon\|^2\right) < \infty$,*

$$\hat{W}_n \overset{P}{\to} W^0$$

Proof Remark that it exist a constant B such that

$$sup_{W \in \Theta_s} \|Y - F_W(Z)\|^2 < \|Y\|^2 + B$$

because Θ_s is compact, so $F_W(Z)$ is bounded. For a matrix $A \in \mathbb{R}^{d \times d}$, let $\|A\|$ be a norm, for example $\|A\|^2 = tr\left(AA^T\right)$. We have

$$\liminf_{W \in \Theta_s} \|\Gamma_n(W)\| = \|\Gamma(W^0)\| > 0$$
$$\limsup_{W \in \Theta_s} \|\Gamma_n(W)\| := C < \infty$$

and since the function :

$$\Gamma \mapsto \ln \det \Gamma, \text{ for } C \geq \|\Gamma\| \geq \|\Gamma(W^0)\|$$

is uniformly continuous, by the same argument that example 19.8 of [Van der Vaart, 1998] the set of functions $U_n(W)$, $W \in \Theta_s$ is Glivenko-Cantelli.

Finally, the theorem 5.7 of [Van der Vaart, 1998], show that $\hat{W}_n$ converge in probability to W^0 ∎.

2.2   Asymptotic normality

For this purpose we have to compute the first and the second derivative with respect to the parameters of $U_n(W)$. First, we introduce a notation : if $F_W(X)$ is a d-dimensional parametric function depending of a parameter W, write $\frac{\partial F_W(X)}{\partial W_k}$ (resp. $\frac{\partial^2 F_W(X)}{\partial W_k \partial W_l}$) for the d-dimensional vector of partial derivative (resp. second order partial derivatives) of each component of $F_W(X)$.

First derivatives :   if $\Gamma_n(W)$ is a matrix depending of the parameter vector W, we get from [Magnus and Neudecker, 1988]

$$\frac{\partial}{\partial W_k} \ln \det \left(\Gamma_n(W)\right) = tr\left(\Gamma_n^{-1}(W)\frac{\partial}{\partial W_k}\Gamma_n(W)\right)$$

Hence, if we note

$$A_n(W_k) = \frac{1}{n}\sum_{t=1}^{n}\left(-\frac{\partial F_W(z_t)}{\partial W_k}(y_t - F_W(z_t))^T\right)$$

using the fact

$$tr\left(\Gamma_n^{-1}(W)A_n(W_k)\right) = tr\left(A_n^T(W_k)\Gamma_n^{-1}(W)\right) = tr\left(\Gamma_n^{-1}(W)A_n^T(W_k)\right)$$

we get
$$\frac{\partial}{\partial W_k}\ln \det\left(\Gamma_n(W)\right) = 2tr\left(\Gamma_n^{-1}(W)A_n(W_k)\right) \tag{3}$$

Second derivatives :   We write now

$$B_n(W_k, W_l) := \frac{1}{n}\sum_{t=1}^{n}\left(\frac{\partial F_W(z_t)}{\partial W_k}\frac{\partial F_W(z_t)}{\partial W_l}^T\right)$$

and

$$C_n(W_k, W_l) := \frac{1}{n} \sum_{t=1}^{n} \left(-(y_t - F_W(z_t)) \frac{\partial^2 F_W(z_t)}{\partial W_k \partial W_l}^T \right)$$

We get

$$\frac{\partial^2 U_n(W)}{\partial W_k \partial W_l} = \frac{\partial}{\partial W_l} 2tr \left(\Gamma_n^{-1}(W) A_n(W_k) \right) =$$
$$2tr \left(\frac{\partial \Gamma_n^{-1}(W)}{\partial W_l} A_n(W_k) \right) + 2tr \left(\Gamma_n^{-1}(W) B_n(W_k, W_l) \right) + 2tr \left(\Gamma_n(W)^{-1} C_n(W_k, W_l) \right)$$

Now, [Magnus and Neudecker, 1988], give an analytic form of the derivative of an inverse matrix, so we get

$$\frac{\partial^2 U_n(W)}{\partial W_k \partial W_l} = 2tr \left(\Gamma_n^{-1}(W) \left(A_n(W_k) + A_n^T(W_k) \right) \Gamma_n^{-1}(W) A_n(W_k) \right) +$$
$$2tr \left(\Gamma_n^{-1}(W) B_n(W_k, W_l) \right) + 2tr \left(\Gamma_n^{-1}(W) C_n(W_k, W_l) \right)$$

so

$$\frac{\partial^2 U_n(W)}{\partial W_k \partial W_l} = 4tr \left(\Gamma_n^{-1}(W) A_n(W_k) \Gamma_n^{-1}(W) A_n(W_k) \right)$$
$$+2tr \left(\Gamma_n^{-1}(W) B_n(W_k, W_l) \right) + 2tr \left(\Gamma_n^{-1}(W) C_n(W_k, W_l) \right) \tag{4}$$

Asymptotic distribution of $\hat{W}_n$: The previous equations allow us to give the asymptotic properties of the estimator minimizing the cost function $U_n(W)$, namely from equation (3) and (4) we can compute the asymptotic properties of the first and the second derivatives of $U_n(W)$. If the variable Z has a moment of order at least 3 then we get the following lemma :

Theorem 2 *Assume that* $E\left(\|\varepsilon\|^2 \right) < \infty$ *and* $E\left(\|Z\|^3 \right) < \infty$, *let* $\Delta U_n(W^0)$ *be the gradient vector of* $U_n(W)$ *at* W^0 *and* $HU_n(W^0)$ *be the Hessian matrix of* $U_n(W)$ *at* W^0.

Write finally

$$B(W_k, W_l) := \frac{\partial F_W(Z)}{\partial W_k} \frac{\partial F_W(Z)}{\partial W_l}^T$$

We get then

1. $HU_n(W^0) \overset{a.s.}{\to} 2I_0$
2. $\sqrt{n} \Delta U_n(W^0) \overset{Law}{\to} \mathcal{N}(0, 4I_0)$
3. $\sqrt{n} \left(\hat{W}_n - W^0 \right) \overset{Law}{\to} \mathcal{N}(0, I_0^{-1})$

where, the component (k, l) *of the matrix* I_0 *is :*

$$tr \left(\Gamma_0^{-1} E \left(B(W_k^0, W_l^0) \right) \right)$$

proof : We can show easily that, for all $x \in \mathbb{R}^d$, we have :

$$\|\frac{\partial F_W(Z)}{\partial W_k}\| \le Cte(1 + \|Z\|)$$
$$\|\frac{\partial^2 F_W(Z)}{\partial W_k \partial W_l}\| \le Cte(1 + \|Z\|^2)$$
$$\|\frac{\partial^2 F_W(Z)}{\partial W_k \partial W_l} - \frac{\partial^2 F_W^0(Z)}{\partial W_k \partial W_l}\| \le Cte\|W - W^0\|(1 + \|Z\|^3)$$

Write

$$A(W_k) = \left(-\frac{\partial F_W(Z)}{\partial W_k}(Y - F_W(Z))^T \right)$$

and $U(W) := \log \det(Y - F_W(Z))$.

Note that the component (k, l) of the matrix $4I_0$ is:

$$E\left(\frac{\partial U(W^0)}{\partial W_k} \frac{\partial U(W^0)}{\partial W_l^0} \right) = E\left(2tr\left(\Gamma_0^{-1} A^T(W_k^0) \right) \times 2tr\left(\Gamma_0^{-1} A(W_l^0) \right) \right)$$

and, since the trace of the product is invariant by circular permutation,

$$E\left(\frac{\partial U(W^0)}{\partial W_k} \frac{\partial U(W^0)}{\partial W_l^0} \right) =$$
$$4E\left(-\frac{\partial F_{W^0}(Z)^T}{\partial W_k} \Gamma_0^{-1}(Y - F_{W^0}(Z))(Y - F_{W^0}(Z))^T \Gamma_0^{-1}\left(-\frac{\partial F_{W^0}(Z))}{\partial W_l} \right) \right)$$
$$= 4E\left(\frac{\partial F_{W^0}(Z)^T}{\partial W_k} \Gamma_0^{-1} \frac{\partial F_{W^0}(Z)}{\partial W_l} \right)$$
$$= 4tr\left(\Gamma_0^{-1} E\left(\frac{\partial F_{W^0}(Z)}{\partial W_k} \frac{\partial F_{W^0}(Z)^T}{\partial W_l} \right) \right)$$
$$= 4tr\left(\Gamma_0^{-1} E\left(B(W_k^0, W_l^0) \right) \right)$$

Now, the derivative $\frac{\partial F_W(Z)}{\partial W_k}$ is square integrable, so $\Delta U_n(W^0)$ fulfills Lindeberg's condition (see [Hall *et al.*, 2001]) and

$$\sqrt{n}\Delta U_n(W^0) \overset{Law}{\to} \mathcal{N}(0, 4I_0)$$

For the component (k, l) of the expectation of the Hessian matrix, remark first that

$$\lim_{n\to\infty} tr\left(\Gamma_n^{-1}(W^0) A_n(W_k^0) \Gamma_n^{-1}(W^0) A_n(W_k^0) \right) = 0$$

and

$$\lim_{n\to\infty} tr\Gamma_n^{-1} C_n(W_k^0, W_l^0) = 0$$

so

$$\lim_{n\to\infty} H_n(W^0) = \lim_{n\to\infty} 4tr\left(\Gamma_n^{-1}(W^0) A_n(W_k^0) \Gamma_n^{-1}(W^0) A_n(W_k^0) \right) +$$
$$2tr\Gamma_n^{-1}(W^0) B_n(W_k^0, W_l^0) + 2tr\Gamma_n^{-1} C_n(W_k^0, W_l^0) =$$
$$= 2tr\left(\Gamma_0^{-1} E\left(B(W_k^0, W_l^0) \right) \right)$$

Now, since $\|\frac{\partial^2 F_W(Z)}{\partial W_k \partial W_l}\| \leq Cte(1 + \|Z\|^2)$ and
$\|\frac{\partial^2 F_W(Z)}{\partial W_k \partial W_l} - \frac{\partial^2 F_W^0(Z)}{\partial W_k \partial W_l}\| \leq Cte\|W - W^0\|(1 + \|Z\|^3)$, by standard arguments found, for example, in [Yao, 2000] we get

$$\sqrt{n}\left(\hat{W}_n - W^0 \right) \overset{Law}{\to} \mathcal{N}(0, I_0^{-1})$$

∎

2.3   Asymptotic distribution of T_n

In this section, we write $\hat{W}_n = \arg\min_{W \in \Theta_s} U_n(W)$ and
$\hat{W}_n^0 = \arg\min_{W \in \Theta_q} U_n(W)$, where Θ_q is view as a subset of $\mathbb{R}^s$. The asymptotic distribution of T_n is then a consequence of the previous section, namely, if we have to replace $nU_n(W)$ by its Taylor expansion around $\hat{W}_n$ and $\hat{W}_n^0$, following [Van der Vaart, 1998] chapter 16 we have :

$$T_n = \sqrt{n}\left(\hat{W}_n - \hat{W}_n^0\right)^T I_0 \sqrt{n}\left(\hat{W}_n - \hat{W}_n^0\right) + o_P(1) \xrightarrow{\mathcal{D}} \chi^2_{s-q}$$

3   Conclusion

It has been show that, in the case of multidimensional output, the cost function $U_n(W)$ leads to a test for the number of parameters in MLP simpler than with the traditional mean square cost function. In fact the estimator $\hat{W}_n$ is also more efficient than the least square estimator (see [Rynkiewicz, 2003]). We can also remark that $U_n(W)$ matches with twice the "concentrated Gaussian log-likelihood" but we have to emphasize, that its nice asymptotic properties need only moment condition on ε and Z, so it works even if the distribution of the noise is not Gaussian. An other solution could be to use an approximation of the covariance error matrix to compute generalized least square estimator :

$$\frac{1}{n}\sum_{t=1}^{n}\left(Y_t - F_W\left(Z_t\right)\right)^T \Gamma^{-1}\left(Y_t - F_W\left(Z_t\right)\right),$$

assuming that Γ is a good approximation of the true covariance matrix of the noise $\Gamma(W^0)$. However it take time to compute a good the matrix Γ and if we try to compute the best matrix Γ with the data, it leads to the cost function $U_n(W)$ (see for example [Gallant, 1987]).

Finally, as we see in this paper, the computation of the derivatives of $U_n(W)$ is easy, so we can use the effective differential optimization techniques to estimate $\hat{W}_n$ and numerical examples can be found in [Rynkiewicz, 2003].

References

[Fukumizu, 2003]K. Fukumizu. Likelihood ratio of unidentifiable models and multilayer neural networks. *Annals of Statistics*, 31:3:533–851, 2003.

[Gallant, 1987]R.A. Gallant. *Non linear statistical models.* J. Wiley and Sons, New-York, 1987.

[Hall and Heyde, 1980]P. Hall and C. Heyde. *Martingale limit theory and its applications.* Academic Press, New-York, 1980.

[Magnus and Neudecker, 1988]Jan R. Magnus and Heinz Neudecker. *Matrix differential calculus with applications in statistics and econometrics.* J. Wiley and Sons, New-York, 1988.

[Rynkiewicz, 2003]J. Rynkiewicz. Estimation of multidimensional regression model with multilayer perceptrons. In J. Mira and J.R. Alvarez, editors, *Computational methods in neural modeling*, volume 2686 of *Lectures notes in computer science*, pages 310–317, 2003.

[Sussman, 1992]H.J. Sussman. Uniqueness of the weights for minimal feedforward nets with a given input-output map. *Neural Networks*, pages 589–593, 1992.

[Van der Vaart, 1998]A.W. Van der Vaart. *Asymptotic statistics*. Cambridge University Press, Cambridge, UK, 1998.

[Yao, 2000]J. Yao. On least square estimation for stable nonlinear ar processes. *The Annals of Institut of Mathematical Statistics*, 52:316–331, 2000.