Efficient Estimation of Multidimensional Regression Model with Multilayer Perceptron

Joseph Rynkiewicz

SAMOS, Universit Paris I - Panthon Sorbonne
Paris, France
rynkiewi@univ-paris1.fr

December 12, 2004

Abstract

This work concerns estimation of multidimensional nonlinear regression models using multilayer perceptron (MLP). The main problem with such models is that we have to know the covariance matrix of the noise to get optimal estimator. however we show that, if we choose as cost function the determinant of the empirical error covariance matrix, or more precisely the logarithm of this determinant, we get an asymptotically optimal estimator.

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). Moreover 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 MLP with parameters or weights W^0.
- (ε_t) is an i.i.d. centered noise with unknown invertible covariance matrix Γ_0.

Our goal is to estimate the true parameter by minimizing an appropriate cost function. This model is called a regression model and a popular choice for the associated cost function is the mean squares error :

$$\frac{1}{n}\sum_{t=1}^{n} \|Y_t - F_W(Z_t)\|^2$$

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

where $\|.\|$ denote the Euclidean norm on $\mathbb{R}^d$. However, it is easy to show that we get then a suboptimal estimator. An other solution is to use an approximation of the covariance error matrix to compute generalized least square estimator :

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

where T denotes the transposition of the matrix and assuming that Γ is a good approximation of the true covariance matrix of the noise Γ_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 asymptotically to the cost function used in this article (see for example Rynkiewicz [4]). Hence, we propose to minimize the cost function

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

This paper is devoted to the theoretical study of this cost function. We assume that the true architecture of the MLP is known so that the Hessian matrix computed in the sequel verifies the assumption to be definite positive (see Fukumizu [1]).

In this framework, we study the asymptotic behavior $\hat{W}_n := \arg\min U_n\left(W\right)$, the weights minimizing the cost function $U_n\left(W\right)$. We show that under simple assumptions this estimator is asymptotically optimal in the sense that it has the same asymptotic behavior than the generalized least square estimator using the true covariance matrix of the noise.

Numerical procedure to compute this estimator and examples of it behavior can be found in Rynkiewicz [4].

2 The first and second derivatives of $W \longmapsto U_n\left(W\right)$

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 $F_W(X)$.

2.1 First derivatives

Now, if $\Gamma_n(W)$ is a matrix depending of the parameter vector W, we get From Magnus and Neudecker [3]

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

here

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

so, these matrix $\Gamma_n(W)$ and it inverse are symmetric.

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

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

2.2 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

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

Now, Magnus and Neudecker [3], 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

$$\begin{array}{l}\frac{\partial^2 U_n(W)}{\partial W_k \partial W_l} = 4tr\left(\Gamma_n^{-1}(W)A_n(W_k)\Gamma_n^{-1}(W)A_n(W_k)\right) \\ +2tr\left(\Gamma_n^{-1}(W)B_n(W_k, W_l)\right) + 2tr\left(\Gamma_n^{-1}(W)C_n(W_k, W_l)\right)\end{array} \tag{3}$$

3 Asymptotic properties 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 (2) and (3) we can compute the asymptotic properties of the first and the second derivatives of $U_n(W)$. Finally, under the assumption that the noise of the model has a moment of order 2, we know that the estimator is strongly consistent (i.e. $\hat{W}_n \overset{a.s.}{\to} W^0$) and, if the noise has a moment of order at least 6 (see the justification in Yao [5]), we get the following lemma :

Lemma 1 *Write W^0 for the true parameter of the model. We recall that $\Gamma_0^{-1} := \Gamma^{-1}(W^0)$ is the true covariance matrix of the noise.*

Let $\Delta U_n(W^0)$ be the gradient vector of $U_n(W)$ at W^0, $\Delta U(W^0)$ be the gradient vector of $U(W) := \log\det(Y - F_W(Z))$ at W^0 and $HU_n(W^0)$ be the Hessian matrix of $U_n(W)$ at W^0.

Write finally

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

and

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

We get then

1. $HU_n(W^0) \stackrel{a.s.}{\to} 2I_0$
2. $\sqrt{n}\Delta U_n(W^0) \stackrel{Law}{\to} \mathcal{N}(0, 4I_0)$

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 To prove the lemma, just 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,

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

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

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

From a classical argument of local asymptotic normality (see for example Yao [5]), we deduce from this lemma the following property for the estimator $\hat{W}_n$:

Proposition 1 *Let W_n^* the estimator of the generalized least square :*

$$W_n^* := \arg\min \frac{1}{n} \sum_{t=1}^{n} \left(Y_t - F_W\left(Z_t\right)\right)^T \Gamma_0^{-1} \left(Y_t - F_W\left(Z_t\right)\right)$$

then we have

$$\lim_{n\to\infty} \sqrt{n}(W_n^* - W^0) = \lim_{n\to\infty} \sqrt{n}(\hat{W}_n - W^0) = \mathcal{N}(0, I_0^{-1})$$

We can see that $\hat{W}_n$ has the same asymptotic behavior than the estimator generalized least square estimator with the true covariance matrix Γ_0^{-1} which is asymptotically optimal (see for example Ljung [2]), so the proposed estimator is asymptotically optimal too.

4 Conclusion

In the linear multidimensional regression model the optimal estimator has an analytic solution (see Magnus and Neudecker [3]), so it doesn't make sense to consider minimization of a cost function. However, for the non-linear multidimensional regression model the ordinary least square estimator is sub-optimal if the covariance matrix of the noise is not the identity matrix. We can overcome this difficulty by using the cost function $U_n(W)$. The numerical computation and the empirical properties of these estimator have been studied in a previous article (see rynkiewicz [4]). In this paper, we have given a proof of the optimality of the estimator associated with $U_n(W)$. This is then the best choice for estimating multidimensional non-linear regression model with multilayer perceptron.

References

[1] Fukumizu, K.: A regularity condition of the information matrix of a multilayer perceptron network Neural Networks, Vol.9, **5** (1996) 871–879

[2] Ljung, L.: System identification : Theory for the user. Prentice Hall (1999)

[3] Magnus, J., Neudecker, H.: Matrix differential calculus with applications in statistics and econometrics. J. Wiley and Sons (1988)

[4] Rynkiewicz, J. : Estimation of Multidimensional Regression Model with Multilayer Perceptrons. Proc of IWANN'03, Lectures Notes in Computer Science **2686** (2003) 310-317

[5] Yao, J.F.: On least square estimation for stable nonlinear AR processes. The Annals of Institut of Mathematical Statistics **52** (2000) 316-331