Estimation of Multidimensional Regression Model with Multilayer Perceptrons

Joseph Rynkiewicz

Université de Paris I,
SAMOS-MATISSE, 90 rue de tolbiac,
75013 Paris,France

Abstract. This work concerns estimation of multidimensional nonlinear regression models using multilayer perceptron (MLP). For unidimensional data, the ordinary least squares estimator matches with the Gaussian maximum likelihood estimator. However, in the multidimensional case, the Gaussian maximum likelihood estimator minimize the determinant of the empirical error's covariance matrix. This paper is devoted to the study of this estimator using a MLP. In particular, we show how to modify the backpropagation algorithm to minimize such cost function and we give heuristic explanations in favor of the use of such function in the multidimensional case.

1 Introduction

Consider a sequence $(Y_t, Z_t)_{t\in\mathbb{N}}$ of random vectors, where $Y_t \in \mathbb{R}^d$, and $Z_t \in \mathbb{R}^{d'}$ (d and d' are positive integer) verifying

$$Y_t = F_{W_0}(Z_t) + \varepsilon_t \tag{1}$$

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 :

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

where $\|.\|$ denote the Euclidean norm on $\mathbb{R}^d$. The weights minimizing this cost function : the ordinary least squares estimator, had been widely studied and if the observations $(Y_t)_{t\in\mathbb{N}}$ are scalar, this estimator matches with the Gaussian maximum likelihood estimator. However, this is not the case if the observations Y_t are d-dimensional with $d \geq 2$.

Indeed, when F_W is a linear function it is well known that the ordinary least square error is a sub-optimal estimator since the best linear unbiased estimator [1] is

$$\bar{W}_n = \arg\min_W \frac{1}{n} \sum_{t=1}^{n} \left(Y_t - F_W\left(Z_t\right)\right)^T \Gamma_0^{-1} \left(Y_t - F_W\left(Z_t\right)\right) \tag{3}$$

where X^T denote the transpose of vector X and Γ_0^{-1} the inverse of Γ_0. In general, the covariance matrix Γ_0 is unknown and we have to estimate this matrix in order to get a better estimator of the weights. For example, Gallant [2] considers the generalized least squares :

$$G_n\left(W, \Gamma\right) := \frac{1}{n} \sum_{t=1}^{n} \left(Y_t - F_W\left(Z_t\right)\right)^T \Gamma^{-1} \left(Y_t - F_W\left(Z_t\right)\right), \tag{4}$$

assuming that Γ is a good approximation of the true covariance matrix of the noise Γ_0. A possible way to construct a sequence of $(\Gamma_k)_{k \in \mathbb{N}^*}$ yielding a good approximation of Γ_0 is the following : using the ordinary least squares estimator $\hat{W}_n$, the noise covariance can be approximated by

$$\Gamma_1 := \Gamma\left(\hat{W}_n\right) := \frac{1}{n} \sum_{t=1}^{n} (Y_t - F_{\hat{W}_n}(Z_t))(Y_t - F_{\hat{W}_n}(Z_t))^T.$$

then, we can use this new covariance matrix to find a generalized least squares estimate $\hat{W}_n^2$:

$$\hat{W}_n^2 = \arg\min_W \frac{1}{n} \sum_{t=1}^{n} \left(Y_t - F_W\left(Z_t\right)\right)^T \left(\Gamma_1\right)^{-1} \left(Y_t - F_W\left(Z_t\right)\right)$$

and calculate again a new covariance matrix

$$\Gamma_2 := \Gamma\left(\hat{W}_n^2\right) = \frac{1}{n} \sum_{t=1}^{n} (Y_t - F_{\hat{W}_n^2}(Z_t))(Y_t - F_{\hat{W}_n^2}(Z_t))^T.$$

It can be shown that this procedure gives a sequence of parameters

$$\hat{W}_n \to \Gamma_1 \to \hat{W}_n^2 \to \Gamma_2 \to \cdots \tag{5}$$

achieving a local maximum of the Gaussian log-likelihood.

However, if we consider the whole parameter (W, Γ), such procedure will be useless because we can directly maximize the Gaussian log-likelihood by minimizing the logarithm of the determinant of the empirical covariance matrix :

$$T_n\left(W\right) := \log\det\left(\frac{1}{n} \sum_{t=1}^{n} (Y_t - F_W(Z_t))(Y_t - F_W(Z_t))^T\right). \tag{6}$$

[1] This estimator is called BLUE

$T_n(W)$ is called the concentrated Gaussian log-likelihood but, naturally, it can be used even if the noise is non-Gaussian.

This paper is devoted to the study of this cost function in the framework of the MLP models. It is organized as follow :

In the second section, we introduce $W_n^* := \arg\min_W T_n(W)$, the weights minimizing the cost function $T_n(W)$. We show how to construct a numerical algorithm to approximate this estimator thanks a modification of the backpropagation algorithm.

In the third section we give heuristic arguments in favor of the use of this estimator when the covariance matrix of the noise is not the identity

In the fourth section we compare the performance of this estimator with the ordinary least square estimator on a simulated example.

2 Minimization of $T_n(W)$

We introduce first some notations :

1. For a $d \times d$ matrix Γ, let $(\Gamma_{ij})_{1 \leq i,j \leq d}$ be the vector $(\Gamma_{11}, \Gamma_{12}, \cdots, \Gamma_{1d}, \Gamma_{21}, \cdots, \Gamma_{2d}, \Gamma_{31}, \cdots, \Gamma_{dd})$.
2. If X is a multidimensional vector, let $X(i)$ be the ith element.
3. If A is an non singular matrix and A^{-1} it's inverse, let a_{ij}^{-1} be the coefficients of A^{-1}.
4. let $tr(A)$ be the sum of diagonal element of the matrix A.

The observations are the data $(y_t, z_t)_{1 \leq t \leq n}$, and we want estimate the parameters $\hat{W}_n^*$ minimizing $T_n(W) = \log\det\left(\frac{1}{n}\sum_{t=1}^n (y_t - F_W(z_t))(y_t - F_W(z_t))^T\right)$. As usual, we cannot find exact solution to such problem. However, we can get a good approximation of the solution with differential optimization. This involve the calculus of the gradient with respect to the weights of the MLP of the cost function which is performed thanks the following modified backpropagation algorithm.

2.1 Calculus of the derivative of $W \longmapsto T_n(W)$:

If $\Gamma_n(W)$ is a matrix depending of the parameter vector W, we get From Magnus and Neudecker [5]

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

Hence, if $\Gamma_n(W) = \sum_{t=1}^n (y_t - F_W(z_t))(y_t - F_W(z_t))^T$, the derivative of $\ln(\det(\Gamma_n(W)))$ with respect to the weight W_k is :

$$\frac{\partial}{\partial W_k}(\ln(\det(\Gamma_n(W)))) = \left(\Gamma_{ij}^{-1}\right)^T_{1 \leq i,j \leq d}\left(\frac{\Gamma_{ij}}{\partial W_k}\right)_{1 \leq i,j \leq d}$$

with

$$\frac{\partial \Gamma_{ij}}{\partial W_k} =$$

$$\frac{1}{n}\sum_{t=1}^{n}\left[-\frac{\partial F_W(z_t)(i)}{\partial W_k}\times\left(y_t - F_W(z_t)\right)(j) - \frac{\partial F_W(z_t)(j)}{\partial W_k}\left(y_t - F_W(z_t)\right)(i)\right] \quad (7)$$

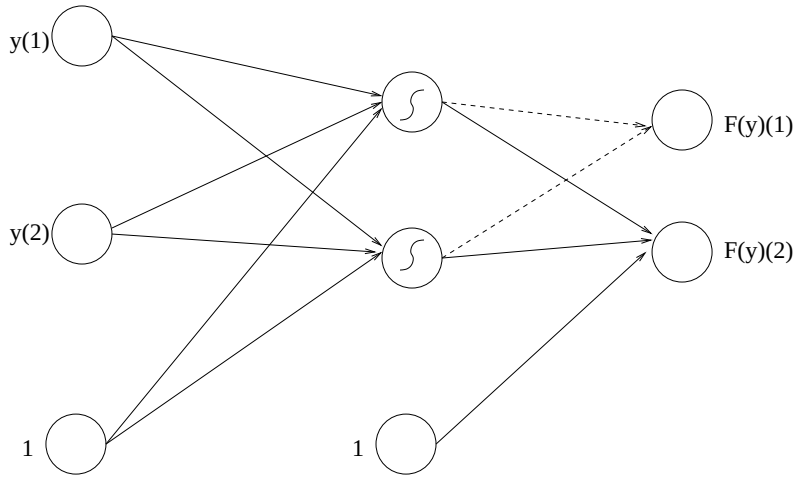
so

$$\frac{\partial}{\partial W_k}(\ln(\det(\Gamma_n(W)))) = \frac{1}{n}\left(\Gamma_{ij}^{-1}\right)_{1\le i,j\le d}^{T}\times$$

$$\left(\sum_{t=1}^{n} -\frac{\partial F_W(z_t)(i)}{\partial W_k}\times\left(y_t - F_W(z_t)\right)(j) - \frac{\partial F_W(z_t)(j)}{\partial W_k}\left(y_t - F_W(z_t)\right)(i)\right)_{1\le i,j\le d}. \quad (8)$$

The quantity $\frac{\partial F_W(z_t)(i)}{\partial W_k}$ is computed by backpropagating the constant 1 for the MLP restricted to the output i. The figure 1 give a example of a MLP restricted to the output 2.

Fig. 1. MLP restricted to the output 2 : the plain lines



Hence, the calculus of the gradient of $T_n(W)$ with respect to the parameters of the MLP is straightforward. We have to compute the derivative with respect to the weights of each single output MLP extracted from our MLP by backpropagating the constant value 1. Then, according to the formula (7), we can easily compute the derivative of each terms of the empirical covariance matrix of the noise. Finally the gradient is obtained by the sum of all the derivative terms of this empirical covariance matrix ponderated by the terms of it's inverse as in formula (8).

2.2 Differential optimization

Using the previous calculus, we can applied one of the numerous techniques of differential optimization to find a local minimum of the cost function $T_n(W)$. We can find a comprehensive review of such techniques in Press et al. [6] and we recommend especially the BFGS algorithm, which is a very fast quasi-newton algorithm.

We note that the calculus of the gradient of $T_n(W)$ is more complex than the calculus of derivatives with the classical mean square criteria, but, in general, optimization algorithms are associated with minimization along a line like Brent's method and the derivatives of the cost function are less evaluated than the function itself. So, finally, the minimization of T_n (W) is only slightly more costly than the minimization of the ordinary mean square criteria.

3 Heuristics on the efficiency of $\hat{W}_n^*$

Under suitable assumptions, see for example Sussmann [7], the model admit a theoretical MLP with an optimal parameter W_0 defined up to a permutation of the weights. Under weak conditions the estimator $\hat{W}_n^*$ converges almost surely to the true parameter, see for example Gourieroux et al. [3]. This result of consistency holds also for the ordinary least square estimator $\hat{W}_n$, see for example White [8].

The differences between $\hat{W}_n^*$ and $\hat{W}_n$ is the speed of convergence or more precisely the variance of the two estimator in function of the number of observation. The better estimator is the estimator with the smallest variance.

First of all, we have to remark that, if the density of the noise is really Gaussian, the estimator $\hat{W}_n^*$ is asymptotically efficient as maximum likelihood estimator. This properties imply that no other consistent estimator can achieve a better variance asymptotically. Gaussian assumption of the noise can appear to be a strong assumption but it is justified by the theorem central limit and the fact that the noise is generally the sum of a lot of random effects.

Moreover, as we have seen in the introduction, maximum Gaussian likelihood estimator matches with the limit estimator of the iterated generalized least square (cf equation (5)). So it matches with the generalized least square estimator with the best approximation of Γ_0 we can get from the observations $(y_t, z_t)_{1 \leq t \leq n}$.

If the regression function is linear the generalized least square estimator using Γ_0 is optimal (BLUE), in the non linear case we have the same property but only asymptotically.

Indeed, if $\psi(W_0)$ is the Jacobian matrix of the MLP function with respect to the true weights, E is the expectation with respect to the true distribution of the data (Y, Z) and $\hat{W}_n^{\Gamma}$ are the weights minimizing cost function $G_n(W, \Gamma)$ (see equation (4)), we get from Yao [9] :

$$\lim_{n \to \infty} \sqrt{n} \left(\hat{W}_n^{\Gamma} - W_0 \right) \stackrel{n \to \infty}{\longrightarrow} \mathcal{N}(0, \Phi_{\Gamma})$$

with

$$\Phi_\Gamma = \left[E\left(\psi\left(W_0\right)\Gamma^{-1}\psi^T\left(W_0\right)\right)\right]^{-1} \times \left[E\left(\psi\left(W_0\right)\Gamma^{-1}\Gamma_0\Gamma^{-1}\psi^T\left(W_0\right)\right)\right] \times \left[E\left(\psi\left(W_0\right)\Gamma^{-1}\psi^T\left(W_0\right)\right)\right]^{-1}$$

Now it is straightforward to establish (see for example Ljung [4]), that

$$\forall \Gamma \ : \ \Phi_\Gamma \geq \Phi_{\Gamma_0}$$

where the inequality is the standard inequality for definite positive matrices[2].

So the less asymptotically variant estimator is obtained when we use the true covariance of the noise to compute the generalized least square criterion. It seems finally natural to use the best estimation of the this true covariance matrix that we can achieve from the data, and so to use the cost function $T_n\left(W\right)$.

4 Simulated example

Although the estimator associated to the cost function $T_n\left(W\right)$, is theoretically better than the ordinary mean least square estimator we have to confirm this fact on simulation. Indeed, their are some pitfalls in practical situations with MLP.

The first point is that we have no guaranty to reach the global minimum of the cost function, we can only hope to find a good local minimum especially if we are using many estimations with different initials weights.

The second point, is the fact that MLP are black box, it means that it is difficult to give an interpretation of their parameters and it is almost impossible to compare MLP by comparing their parameters even if we try to take into account the possible permutations of the weights.

All these reasons explain why we choose to compare the estimated covariance matrices of the noise instead of compare directly the estimated parameters of MLP.

4.1 The model

To simulate our data, we use a MLP with 2 inputs, 3 hidden units, and 2 outputs. We choose to simulate a time series because it is very easy task as the outputs at time t are the inputs for the time $t+1$. Moreover, with MLP, the statistical properties of such model are the same than with independent identically distributed (i.i.d.) data.

The equation of the model is the following

$$Y_{t+1} = F_{W_0}\left(Y_t\right) + \varepsilon_{t+1}$$

where

[2] We say $\Phi_\Gamma \geq \Phi_{\Gamma_0}$ if and only if $\Phi_\Gamma - \Phi_{\Gamma_0}$ is a positive semidefinite matrix

- $Y_0 = (0, 0)$.
- $(Y_t)_{1 \leq t \leq 1000}$, $Y_t \in \mathbb{R}^2$, is the bidimensional simulated random process
- F_{W_0} is a MLP function with weights W_0 chosen randomly between -2 and 2.
- (ε_t) is an i.i.d. centered noise with covariance matrix $\Gamma_0 = \begin{pmatrix} 1.81 & 1.8 \\ 1.8 & 1.81 \end{pmatrix}$.

In order to study empirically the statistical properties of our estimator we make 10 independent simulations of the bidimensional times series of length 1000.

On each time series we estimate the weights of the MLP using the cost function $T_n(W)$ and the ordinary least square estimator (MCO). The estimations have been done using the second order algorithm BFGS, and for each estimation we choose the best results obtained after 20 random initializations of the weights. Doing so, we avoid to plague our learning with poor local minima.

We show here the mean of estimated covariance matrices of the noise for the different estimators:

$$T_n(W) : \begin{pmatrix} 1.793 & 1.785 \\ 1.785 & 1.797 \end{pmatrix} \text{ and } MCO : \begin{pmatrix} 1.779 & 1.767 \\ 1.767 & 1.783 \end{pmatrix}$$

the estimated standard deviation of the terms of the matrices are all equal to 0.003, so the differences observed between the two matrices are statistically significants. We can see that the estimated covariance of the noise is in mean better with the estimator associated to the cost function $T_n(W)$, in particular it seems that there is slightly less overfitting with this estimator, and the non diagonal terms are greater than with the estimator associated with the MCO. Indeed, as expected, the determinant of the mean matrix associated to $T_n(W)$ is 0.036 instead of 0.050 for the matrix associated to the MCO.

5 Conclusion

In the multidimensional case the ordinary least square estimator are often suboptimal if the covariance matrix of the noise is not the identity matrix. In seeking to take into account the covariance matrix of the noise we find that it is natural to use the concentrated log-likelihood as cost function. We have shown that the differential minimization of this cost function is easy with MLP, since we can compute the gradient of this function tanks a modification of the backpropagation algorithm. Finally the theoretical advantages of this estimator have been verified on a simulation and we can expect a amelioration of the learning process in using this cost function. Even if this amelioration is small, it can be very important to improve the variance of the parameter especially when we are using pruning techniques based on this variance like the SSM algorithm of Cottrell et al. [1].

References

1. Cottrell, M. Girard, B. and Y., Mangeas, M., Muller, C.: Neural Modeling for Time Series : a Statistical Stepwise Method for Weight Elimination. IEEE Trans on Neural Networks **6:6** (1995) 1355–1364
2. Gallant, R.: Non linear statistical models. J. Wiley and Sons (1987)
3. Gourieroux, C., Monfort, A., Trognon, A.: Pseudo maximum likelihood methods: Theory. Econometrica **52:3** (1984) 681–700
4. Ljung, L.: System identification : Theory for the user. Prentice Hall (1999)
5. Magnus, J., Neudecker, H.: Matrix differential calculus with applications in statistics and econometrics. J. Wiley and Sons (1988)
6. Press, W., Flannery, B., Teukolsky, S., Vetterling, W.: Numerical recipes in C : The art of scientific computing. Cambridge University Press (1992)
7. Sussmann, H.: Uniqueness of the wieghts for minimal feedforward nets with a given input-output map. Neural Networks, **5** (1992) 589-593
8. White, H.: Artificial neural networks. Blackwell (1992)
9. Yao, J.F.: On least square estimation for stable nonlinear AR processes. The Annals of Institut of Mathematical Statistics **52** (2000) 316-331