ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 57-64

Parametric Bootstrap for Test of Contrast Difference in Neural Networks

Riadh Kallel, Joseph Rynkiewicz
MATISSE-SAMOS UMR 8595
90, rue de Tolbiac
75634 Paris cedex 13
kallel,rynkiewi@univ-paris1.fr

Abstract

This work concernes the contrast difference test and its asymptotic properties for non linear auto-regressive models. Our approach is based on an application of the parametric bootstrap method. It is a re-sampling method based on the estimate parameters of the models. The resulting methodology is illustrated by simulations of multilayer perceptron models, and an asymptotic justification is given at the end.

1 Introduction

To deal with neural network models, we use statistical theoretical results to compare several models and choose the best. But in many cases, we have not a very large number of data and the theoretical results can lead to an over-parametrized model. In this paper, we recall the asymptotic properties of the minimum contrast estimator (section 3) and some empirical results about the test of contrast difference (section 4). In section 5, we propose a selection method using the parametric bootstrap. This method consists in replacing the asymptotic distribution of the test statistic by the empirical distribution built from the bootstrapped samples. An asymptotic justification is given in section 6.

2 Non Linear Auto-Regressive Models

Let $p, d > 1$ be two positive integers. A functional auto-regressive process on $\mathbb{R}^d$ is a sequence of random vectors defined by:

$$X_t = F_\theta(X_{t-1}, \ldots, X_{t-p}) + \epsilon_t \tag{1}$$

where ϵ_t is an i.i.d. noise with mean 0 and constant variance σ^2, and where function F is known. The parameter θ belongs to a subset Θ of $\mathbb{R}^s$ ($s \in \mathbb{N}^*$). Such a model is denoted below by $ARF_d(p)$.

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 57-64

Let $||.||$ be the Euclidean norm on $\mathbb{R}^d$. We define *the contrast process* associated to the least squares by:

$$U_n(\theta) = \frac{1}{n}\sum_{t=1}^{n} ||X_t - F_\theta(X_{t-1}, \ldots, X_{t-p})||^2, \tag{2}$$

and *the least squares estimator* by:

$$\hat{\theta}_n = \operatorname{Arg}\min_{\theta\in\Theta} U_n(\theta). \tag{3}$$

In the same way, we define *the contrast process* associated to the Gaussian log-likelihood by:

$$V_n(\theta) = \log\det\left(\frac{1}{n}\sum_{t=1}^{n}(X_t - F_\theta(X_{t-1}, \ldots, X_{t-p}))(X_t - F_\theta(X_{t-1}, \ldots, X_{t-p}))^T\right), \tag{4}$$

and *the maximum likelihood estimator* by:

$$\tilde{\theta}_n = \operatorname{Arg}\min_{\theta\in\Theta} V_n(\theta). \tag{5}$$

In this paper, we deal with the asymptotic properties of the estimators $\hat{\theta}_n$ and $\tilde{\theta}_n$ as well as the tests of contrast differences.

3 Asymptotic Properties of the Minimum Contrast Estimator

3.1 Consistency and Asymptotic Normality

In the following, we assume that the regression function is a function which can be implemented by a multi-layer perceptron (MLP), with one hidden layer, hyperbolic tangent activation functions for the hidden units and one linear output unit. Besides, we assume that the process $(X_t)_{1\leq t\leq n}$ is stationary with invariant measure μ. We have the following theorem (see Yao [5]):

Theorem 1 **Strong Consistency and Asymptotic Normality**
Assume that:

1. $(\varepsilon_t)_{t>0}$ *is an i.i.d. sequence, such that* $E(\varepsilon_t^6) < \infty$,
2. θ *belongs to a compact subset* Θ *in the Euclidean space* $\mathbb{R}^s$ *of dimension* s, *such that* $\theta_0 \in \overset{\circ}{\Theta}$ *(interior of* Θ*).*
3. *(Identifiability Condition) For all* θ *different from* θ_0, $F_\theta \neq F_{\theta_0}$ *in the sense that there exists* $\tilde{y} \in \mathbb{R}^p$ *such that* $F_\theta(\tilde{y}) \neq F_{\theta_0}(\tilde{y})$.

4. The $m \times m$ matrix

$$\Sigma_0 = \int_{\mathbb{R}^p} \Big[\frac{\partial}{\partial \theta_i} F_{\theta_0}(\tilde{y}) \frac{\partial}{\partial \theta_j} F_{\theta_0}(\tilde{y}) \Big]_{1 \leq i,j \leq m} \; \mu(d\tilde{y}), \tag{6}$$

is positive definite.

Then

(a) The estimator $\hat{\theta}_n$ is strongly consistent, that is it almost surely converges to θ_0 when n goes to $+\infty$.

(b) The term $\sqrt{n}\left[\hat{\theta}_n - \theta_0\right]$ converges in distribution to the multidimensional Gaussian distribution $\mathcal{N}(0, \sigma^2 \Sigma_0^{-1})$.

If the noise ε is assumed to be Gaussian, a similar theorem can be proved for the estimator $\tilde{\theta}_n$ (see Rynkiewicz [3]).

3.2 Asymptotic Test of Contrast Difference

Let q be an integer less than s. We put $H_s = \{\theta \in \Theta \subset \mathbb{R}^s\}$. A sub-hypothesis H_q of H_s means that θ belongs to a subset of Θ with parametric dimension q less than s.

In the following, Θ^q (resp. Θ^s) is the parameter space of dimension q (resp. s). To test the hypothesis H_q against H_s, we use the contrast difference statistic:

$$T_n = 2n[U_n(\hat{\theta}_n^q) - U_n(\hat{\theta}_n^s)],$$

where $\hat{\theta}_n^q = \text{Arg}\min_{\theta^q \in \Theta^q} U_n(\theta^q)$ (resp. $\hat{\theta}_n^s = \text{Arg}\min_{\theta^s \in \Theta^s} U_n(\theta^s)$). Then the next result defines the asymptotic distribution of statistic T_n used to test hypothesis H_q against hypothesis H_s:

Proposition 1 *Under the assumptions of theorem 1, under hypothesis H_q, we have*

$$T_n \xrightarrow{\mathcal{L}} \sum_{i=1}^{s-q} \lambda_i \chi_i^2,$$

where the χ_i^2 are χ-square independent variables with 1 degree of freedom, and the λ_i are positive scalars. ■

Remark 1 *In the scalar case, $\lambda_1 = \ldots = \lambda_{s-q} = 2\sigma^2$ and the test statistic possess a χ-square distribution with $s - q$ degrees of freedom.*

Remark 2 *In the maximum likelihood case, by assuming that the noise is Gaussian and by defining*

$$S_n = n[V_n(\tilde{\theta}_n^q) - V_n(\tilde{\theta}_n^s)],$$

one has

$$S_n \xrightarrow{\mathcal{L}} \sum_{i=1}^{s-q} \chi_i^2,$$

which is the classical behavior of the likelihood ratio test.

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 57-64

4 Empirical Behavior of the Test Statistics

The goal of this section is to study the asymptotic behavior of the contrast difference test using a simulated example, with a finite number n of observations. Let us define a MLP F_{θ_0} with three inputs, two hidden units and one output. This MLP is chosen as the true model (unknown in the practical case). We run the following experience:

1. simulate 100 independent databases $(B(b))_{1\leq b\leq 100}$ with F_{θ_0} and $\varepsilon \sim \mathcal{N}(0,\sigma^2)$ ($\sigma^2 = 0.25$),
2. on each database, estimate two MLP $F_{\hat{\theta}^q}$ et $F_{\hat{\theta}^s}$, such that $F_{\hat{\theta}^q}$ has the same architecture as à F_{θ_0} (with $q = 11$ parameters) and $F_{\hat{\theta}^s}$ has one hidden unit more (and $s = 16$ parameters),
3. compute the statistics $(T_n(b))_{1\leq b\leq 100}$ and $(S_n(b))_{1\leq b\leq 100}$.

The distributions of the two variables $\frac{1}{2\sigma^2}T_n$ et S_n would possess, approximately if n is large enough, a χ-square distribution with $s-q=5$ degrees of freedom since hypothesis H_q is true. The histograms of $(T_n(b))_{1\leq b\leq 100}$ and $(S_n(b))_{1\leq b\leq 100}$ (Fig.1) show the convergence to the $\chi^2_{(5)}$ distribution when n increases, but the convergence speed is very slow: we recognize the shape of a $\chi^2_{(5)}$ distribution only for $n = 100000$.

If we use the contrast difference test as a criterion to choose between the two models, with a size of type I error (level) of 5%, we reject H_q in 85% of the cases for $n = 1000$, and in 50% of the cases for $n = 10000$.

In neural network applications, we have not always a very large number of data. So applying this method can conduce to accept an over-parameterized multi-layer perceptron if we apply the proposition (1) to select the good model among the different studied perceptrons. In this paper, we propose to apply a re-sampling method based on the so-called "parametric bootstrap" to improve the results.

5 Parametric Bootstrap

5.1 Methodology

The parametric bootstrap (Efron [2], Davison [1], White [4]) method is a re-sampling method based on the estimated parameters of the model. In fact, it would be ideal to be able to use the empirical distributions of section 4 to determine the critical region of our test. However, here, the true parameter is unknown. So we replace this true parameter by its estimate in a model corresponding to hypothesis H_q (null hypothesis).

Let us assume that we want to test hypothesis H_s against null hypothesis H_q with a 5% level. Then we apply the following procedure:

- estimate parameter θ_0 in a q-dimensional model and in a s-dimensional model. Let $\hat{\theta}^q$ and $\hat{\theta}^s$ the estimates that we get. We denote by T_n and S_n the test statistics associated with these estimates,

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 57-64

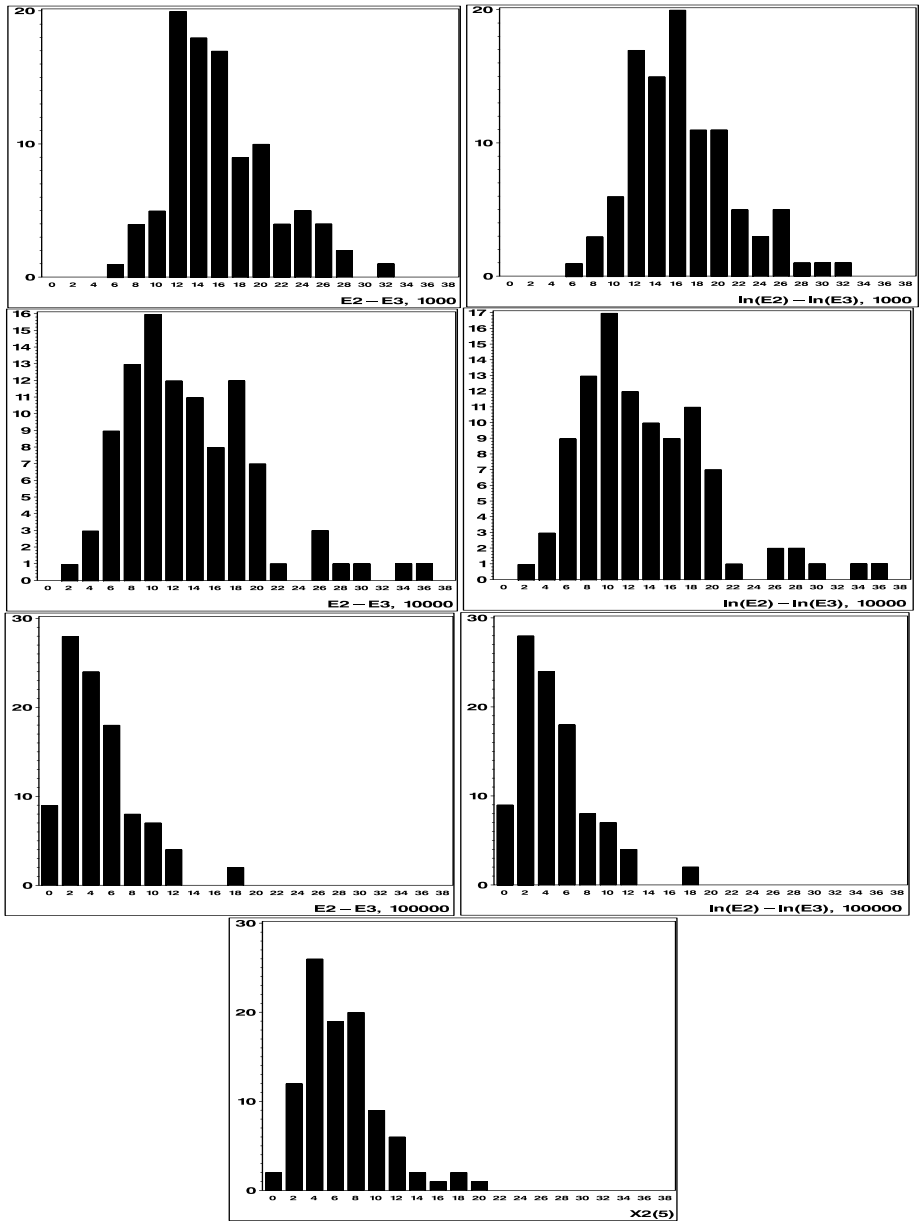

Figure 1: First column: Histogram of variable $(T_n(b))_{1\leq b\leq 100}$ for $n = 1000, n = 10000, n = 100000$. Second column: Histogram of variable $(S_n(b))_{1\leq b\leq 100}$ for the three applications. Last line represents the histogram for a direct simulation of a $\chi^2_{(5)}$ distribution.

- simulate 100 independent databases having the same size n as the initial sample by using the estimate $\hat{\theta}^q$ and $\varepsilon \sim \mathcal{N}(0, \hat{\sigma}^2)$, where $\hat{\sigma}^2$ is the estimated variance associated to $\hat{\theta}^q$,
- on each database b, estimate two MLP $F_{\hat{\theta}^q(b)}$ and $F_{\hat{\theta}^s(b)}$,
- compute the statistics $(T_n(b))_{1 \leq b \leq 100}$ and $(S_n(b))_{1 \leq b \leq 100}$,
- draw the histograms of $(T_n(b))_{1 \leq b \leq 100}$ and $(S_n(b))_{1 \leq b \leq 100}$, and determine the α-quantile of T_n (resp. S_n) : C_α (resp. D_α), which provides the critical region given by $T_n > C_\alpha$ (resp. $S_n > D_\alpha$),
- if T_n (resp. S_n) belongs to the critical region, one rejects the null hypothesis H_q.

The main idea is to replace the asymptotic distribution of the test statistic by the distribution that we construct from the bootstrapped samples.

5.2 Empirical Verification of the Method

We applied the procedure which is described in the previous section to ten models with the same dimension q, but different parameters. The number of observations is fixed, equal to $n = 1000$ and we replace the asymptotic distribution of the contrast difference test (the $\chi^2_{(s-q)}$ distribution) by the empirical distribution. The results are displayed in table 1. If we consider the level 5%, we accept the null hypothesis (H_q) for all the models. The size of type I error is well controlled, what was not true when we used the theoretical asymptotic distribution.

Simulations with H_q	C_α	T_n	D_α	S_n
sim1	23.86	18.98	23.13	18.57
sim2	22.56	15.51	23.40	16.03
sim3	25.53	8.87	26.33	8.98
sim4	21.76	6.73	22.68	7.12
sim5	22.86	16.62	24.43	17.18
sim6	27.45	11.50	29.06	12.19
sim7	22.52	12.56	23.80	12.91
sim8	22.12	16.47	23.10	16.74
sim9	22.17	12.03	22.52	12.19
sim10	25.54	13.62	24.91	13.30

Table 1: α-quantiles and statistics of T_n and S_n with $\alpha = 5\%$ and n=1000. For all cases, we accept the null hypothesis H_q.

5.3 Power of the Test

In this part, we deal with the power of the contrast difference test combined with the parametric bootstrap method. We have to study the behavior of this method when the

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 57-64

alternative (H_s) is true. We did exactly the same computations as previously, but this time, the ten initial data bases were simulated by using MLP with an architecture corresponding to H_s. As in the previous section, we simulate 10 databases, and apply to each of them the re-sampling method described in section 5.1 to estimate the α-quantiles.

The results are presented in Table 2. We can observe that for all the simulations, we reject the null hypothesis H_q, and we select the true model corresponding to H_s.

Simulations with H_s	C_α	T_n	D_α	S_n
sim1	22.07	57.70	21.73	58.26
sim2	25.62	48.83	22.89	45.89
sim3	22.34	66.80	22.77	65.92
sim4	25.70	86.40	24.12	79.84
sim5	20.70	48.60	22.35	52.31
sim6	21.82	44.97	23.16	48.02
sim7	23.57	45.33	22.72	45.93
sim8	21.59	83.52	22.58	83.61
sim9	21.59	69.70	22.57	71.62
sim10	23.50	86.34	22.36	81.64

Table 2: α-quantiles and statistics of T_n and S_n with $\alpha = 5\%$ and n=1000. In all cases, we reject the null hypothesis H_q.

6 Asymptotic Justification of the Method

Although the main usefulness of the proposed method is that it is more accurate than the one based only on the asymptotic distribution, it would be more satisfactory to verify that these methods are justified when the length n of the observed series grows to infinity.

First, we can notice that for the likelihood ratio test, nothing remains to be proven, since the asymptotic distribution for this test does not depend on the parameters, but only on the model dimensions.

For the least squares case, according to proposition (1), one has:

$$T_n = 2n[\overline{U}_n(\hat{\alpha}_n) - U_n(\hat{\theta}_n)] \xrightarrow{C.L} \sum_{i=1}^{s-q} \lambda_i \chi_i^2.$$

Moreover, for $d = 1$, we know the values λ_i since they are all equal to $2\sigma^2$.

So, for each base that we build by re-sampling, we have for all b, $1 \leq b \leq 100$

$$T_n(b) \xrightarrow{C.L} 2\hat{\sigma}_n^2(b) \sum_{i=1}^{s-q} \chi_i^2,$$

since we use the estimate variance $\hat{\sigma}_n^2$ instead of the true variance for the simulations.

ESANN'2002 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 24-26 April 2002, d-side publi., ISBN 2-930307-02-1, pp. 57-64

However, as $\hat{\sigma}_n^2(b) \xrightarrow{p.s} \sigma^2$, a straightforward application of Slutsky's theorem gives:

$$T_n(b) \xrightarrow{C.L} 2\sigma^2 \sum_{i=1}^{s-q} \chi_i^2.$$

7 Conclusion

In this paper, we have investigated the use of parametric bootstrap method for contrast difference test for neural network models. We have proven that this method improves the test for $ARF_d(p)$ models when we have not a very large number of data to use.

We note that this justification is no more valuable for multidimensional observations, since in that case the λ_i depend on the unknown distribution of the studied process.

References

[1] Davison A. and Hinklay (1996) *Bootstrap methods and their applications*, Combrige University Press.

[2] Efron, B and Tibshirani R. (1993) *An introduction to the bootstrap*, Chapman and Hall.

[3] Rynkiewicz J. (2000) *Modèles hybrides intégrant des réseaux de neurones artificiels à des modèles de chaines de Markov cachées : Application à des prédictions de séries temporelles.* Thèse, Université Paris 1.

[4] White H. and Racine J. (2001) Statistical Inference, The Bootstrap, and Neural-Network Modeling With Application to Foreign Exchange Rates, *IEEE Transactions on Neural Networks*, **12**, n°4, p 657–673.

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