ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Quantile regression with multilayer perceptrons.

S.-F. Dimby and J. Rynkiewicz

Universite Paris 1 - SAMM
90 Rue de Tolbiac, 75013 Paris - France

Abstract. We consider nonlinear quantile regression involving multilayer perceptrons (MLP). In this paper we investigate the asymptotic behavior of quantile regression in a general framework. First by allowing possibly non-identifiable regression models like MLP's with redundant hidden units, then by relaxing the conditions on the density of the noise. In this paper, we present an universal bound for the overfitting of such model under weak assumptions. The main application of this bound is to give a hint about determining the true architecture of the MLP quantile regression model. As an illustration, we use this theoretical result to propose and compare effective criteria to find the true architecture of such regression model.

1 Introduction

Quantiles are points taken at regular intervals from the cumulative distribution function (CDF) of a random variable. Some q-quantiles have special names : The 2-quantile is called the median, the 4-quantiles are called quartiles and the 10-quantiles are called deciles.

We can define the quantile through a simple alternative expedient as an optimization problem. Just as we can define the sample means as the solution to the problem of minimizing a sum of squared residuals, we can define the median as the solution to the problem of minimizing a sum of absolute residuals. More generally, if $y_1, \cdots, y_n$ are observed values, solving

$$\min_{m \in \mathbb{R}} \sum_{i=1}^{n} \rho_\tau(y_i - m) \tag{1}$$

where the cost function $\rho_\tau(z) = \tau \times (z) \times \mathbf{1}_{\mathbb{R}^+}(z) - (1-\tau) \times z \times \mathbf{1}_{\mathbb{R}^-}(z)$ is the tilted absolute function. Having succeeded in defining the unconditional quantiles as an optimization problem, it is easy to define conditional quantiles in an analogous fashion. To obtain an estimate of the conditional quantile, we simply replace the scalar m in the equation 1 by a function $f(x_i)$, where x_i are the covariate variables.

2 The model

The basic model is a possibly nonlinear regression model with an additive error. It is given by

$$Y_t = f_\theta(X_t) + \varepsilon_t \tag{2}$$

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Where $(Y_t)_{1\leq t\leq n}$ are the observations, $(X_t)_{1\leq t\leq n}$ are random covariates and $(\varepsilon_t)_{1\leq t\leq n}$ are unobserved error term. The regression function f_θ is assumed to be an MLP function with k hidden units can be written :

$$f_\theta(x) = \beta + \sum_{i=1}^{k} a_i \phi\left(w_i^T x + b_i\right),$$

with $\theta = (\beta, a_1, \cdots, a_k, b_1, \cdots, b_k, w_{11}, \cdots, w_{1d}, \cdots, w_{k1}, \cdots, w_{kd})$ the parameter vector of the model and ϕ a bounded transfer function, usually a sigmoïdal function. θ belongs to $\Theta_k \subset \mathbb{R}^{k\times(d+2)+1}$, a compact (i.e. closed and bounded) set of possible parameters. The quantile regression estimator $f_{\hat{\theta}_\tau}$ is obtained by solving the optimization problem :

$$\begin{array}{l} \min_{\theta\in\Theta_k} M_n^\tau(f_\theta) \\ \text{with } M_n^\tau(f_\theta) = \sum_{i=1}^n \rho_\tau(y_i - f_\theta(x_i)) \end{array} \tag{3}$$

For a function $\rho_\tau(.)$ equal to

$$\rho_\tau(z) = \tau \times (z) \times \mathbf{1}_{\mathbb{R}^+}(z) - (1-\tau) \times z \times \mathbf{1}_{\mathbb{R}^-}(z) \tag{4}$$

In the sequel, let f_{θ_τ} be a, possibly not unique, function such that

$$f_{\theta_\tau} = \arg\min_{\theta\in\Theta_k} M(f_\theta) \text{ with } M(f_\theta) = \int \rho_\tau(y - f_\theta(x)) dP(x,y). \tag{5}$$

f_{θ_τ} is the optimal function for the theoretical quantile regression problem.

2.1 Asymptotic distribution

If the possible functions f_θ are parametric, identifiable and smooth enough function and if the density of the noise exists and is positive then asymptotic normality of the M-estimator can be shown (see Koenker and Basset [1] for the linear case and Weiss [6] for the non-linear case and $\frac{1}{2}$-quantile). However it is possible to give more general results using empirical processes theory. In this paper we prove a general bound valid even if the optimal functions f_{θ_τ} are not unique and without assumptions on the density of noise, except moment conditions.

2.1.1 A general bound for $M_n^\tau(f_\theta)$

We will prove an inequality bounding the difference:

$$M_n^\tau(f_\theta) - M_n^\tau(f_{\theta_\tau}).$$

For an square integrable function $g(X,Y)$ the L_2 norm is:

$$\|g(X,Y)\|_2 := \sqrt{\int g^2(x,y) dP(x,y)}.$$

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Let $\lambda > 0$ be a constant, the generalized derivative function is defined as:

$$\begin{aligned} d_\theta^\lambda(X,Y) &= \frac{\frac{e^{-\lambda\rho_\tau(Y-f_\theta(X))}-e^{-\lambda\rho_\tau(Y-f_{\theta_\tau}(X))}}{e^{-\lambda\rho_\tau(Y-f_{\theta_\tau}(X))}}}{\|\frac{e^{-\lambda\rho_\tau(Y-f_\theta(X))}-e^{-\lambda\rho_\tau(Y-f_{\theta_\tau}(X))}}{e^{-\lambda\rho_\tau(Y-f_{\theta_\tau}(X))}}\|_2} \\ &= \frac{e^{-\lambda\rho_\tau(Y-f_\theta(X))-\lambda\rho_\tau(Y-f_{\theta_\tau}(X))}-1}{\|e^{-\lambda\rho_\tau(Y-f_\theta(X))-\lambda\rho_\tau(Y-f_{\theta_\tau}(X))}-1\|_2} \end{aligned} \tag{6}$$

and let us define $\left(d_\theta^\lambda\right)_-(x,y) = \min\left\{0, d_\theta^\lambda(x,y)\right\}$. For now, let us assume that d_θ^λ is well defined, this point will be discuss later. We can state the following inequality:

Inequality:
for $\lambda > 0$,

$$\sup_{\theta\in\Theta_k} \times (M_n^\tau(f_{\theta_\tau}) - M_n^\tau(f_\theta)) \leq \frac{1}{2\lambda} \sup_{\theta\in\Theta_k} \frac{\sum_{i=1}^n d_\theta^\lambda(x_i,y_i)}{\sum_{i=1}^n \left(d_\theta^\lambda\right)_-^2(x_i,y_i)} \tag{7}$$

Proof:
The proof is very similar to the proof for the least square estimator obtained by Rynkiewicz [4]. We have

$$\begin{aligned} &(M_n^\tau(f_{\theta_\tau}) - M_n^\tau(f_\theta)) = \\ &\tfrac{1}{\lambda}\textstyle\sum_{i=1}^n \log\left(1 + \|\frac{e^{-\lambda\rho_\tau(Y-f_\theta(X))}-e^{-\lambda\rho_\tau(Y-f_{\theta_\tau}(X))}}{e^{-\lambda\rho_\tau(Y-f_{\theta_\tau}(X))}}\|_2 d_\theta^\lambda(x_i,y_i)\right) \\ &\leq \sup_{0\leq p\leq \|\frac{e^{-\lambda\rho_\tau(Y-f_\theta(X))}-e^{-\lambda\rho_\tau(Y-f_{\theta_\tau}(X))}}{e^{-\lambda\rho_\tau(Y-f_{\theta_\tau}(X))}}\|_2} \tfrac{1}{\lambda}\textstyle\sum_{i=1}^n \log\left(1 + p d_\theta^\lambda(x_i,y_i)\right) \\ &\leq \sup_{p\geq 0} \tfrac{1}{\lambda}\left(p\textstyle\sum_{i=1}^n d_\theta^\lambda(x_i,y_i) - \frac{p^2}{2}\sum_{i=1}^n \left(d_\theta^\lambda\right)_-^2(x_i,y_i)\right). \end{aligned}$$

Since for any real number u, $\log(1+u) \leq u - \frac{1}{2}u_-^2$. Finally, replacing p by the optimal value, we found

$$(M_n^\tau(f_{\theta_\tau}) - M_n^\tau(f_\theta)) \leq \tfrac{1}{2\lambda}\frac{\sum_{i=1}^n d_\theta^\lambda(x_i,y_i)}{\sum_{i=1}^n \left(d_\theta^\lambda\right)_-^2(x_i,y_i)}$$

■

This inequality allows to prove that $M_n^\tau(f_{\theta_\tau}) - M_n^\tau(f_\theta)$ is bounded in probability under simple assumptions. This may be applied to model selection as discussed in the next section.

2.2 Application : selection of models

In this section, the set Θ of possible parameters will be set to

$$\Theta = \cup_{k=1}^K \Theta_k,$$

with $\Theta_{k_1} \subset \Theta_{k_2}$ for $k_1 < k_2$ and K is a, possibly huge, fixed constant. Let k^0 be the minimal dimension of the functional space needed to realize the true regression function f_τ. For multilayer perceptron Θ_k may be set of MLP with k hidden units. We define the minimum-penalized estimator of k^0, as the minimizer $\hat{k}$ of

$$T_n(k) = \min_{\theta\in\Theta}\left(M_n^\tau(f_\theta) + a_n(k)\right) \tag{8}$$

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Let us assume the following assumptions:

(A1) $a_n(.)$ is increasing, $n \times (a_n(k_1) - a_n(k_2))$ tends to infinity as n tends to infinity, for any $k_1 > k_2$ and $a_n(k)$ tends to 0 as n tends to infinity for any k.

(A2) It exists $\lambda > 0$ so that $\left\{d_\theta^\lambda, \theta \in \Theta\right\}$ is a Donsker class (see van der Vaart [5]).

We now have:
Theorem:
Under **(A1)** *and* **(A2)**, $\hat{k}$ *converges in probability to the true dimension* k^0.

The proof of this theorem is exactly the same as in Rynkiewicz [4].

The assumption **(A1)** is fairly standard for model selection, in the Gaussian case **(A1)** will be fulfilled by BIC-like criteria. The assumption **(A2)** is more difficult to check. First we note:

$$\begin{array}{l}\left(e^{-\lambda(\rho_\tau(Y-f_\theta(X))-\rho_\tau(Y-f_{\theta_\tau}(X)))}-1\right)^2 = \\ e^{-2\lambda(\rho_\tau(Y-f_\theta(X))-\rho_\tau(Y-f_{\theta_\tau}(X)))} - 2e^{-\lambda(\rho_\tau(Y-f_\theta(X))-\rho_\tau(Y-f_{\theta_\tau}(X)))}+1\end{array}$$

So, d_θ^λ is well defined if $E\left[e^{-2\lambda(\rho_\tau(Y-f_\theta(X))-\rho_\tau(Y-f_{\theta_\tau}(X)))}\right] < \infty$, Since an MLP function is bounded, d_θ^λ is well defined if Y admits exponential moments. Finally, using the same techniques of reparameterization as in Rynkiewicz [3], assumption **(A2)** can be shown to be true for linear regressions or MLP models with sigmoïdal transfer functions, if the set of possible parameters Θ is compact.

3 A little experiment

The theoretical penalization terms of the previous section can be chosen among a wide range of functions (see condition **A1**). In the sequel, a little experiment is conducted to assess the right rate of penalization to guess the "true" architecture of a model.

Consider a simulated model:

$$Y_t = F_{\theta^0}(X_{1t}, X_{2t}) + \varepsilon_t, t = 1, \cdots, n,$$

with $((X_{11}, X_{21}), \cdots, (X_{1n}, X_{2n}))$ i.i.d., $(X_{1t}, X_{2t}) \sim \mathcal{N}(0_{\mathbb{R}^2}, 3 \cdot I_2)$, where I_2 is the identity matrix. The noise sequence $\varepsilon_1, \ldots, \varepsilon_n$ is independent and identically distributed following a Gaussian distribution $\mathcal{N}(0, 1)$ and

$$\begin{array}{l}F_{\theta^0}(x_1, x_2) = \tanh(6 \cdot x_1 - 2 \cdot x_2) + 2 \cdot \tanh(8 - x_1 + 3 \cdot x_2) \\ -3 \cdot tanh(2 - 6 \cdot x_1 - 2 \cdot x_2) + 1.5.\end{array} \quad (9)$$

Here, the true model is an MLP with 2 inputs, 3 hidden units and one output. In order to avoid too long time of computation, the number of hidden units is assumed to be between 1 and 10.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Let D be the size of the parameter vector (the dimension of the model or the number of weights of the MLP), we consider the quantile regression with $\tau = 0.5$, so we minimize the sum of absolute residuals.

We will compare 3 criteria, from the least penalized (AIC like) to the most penalized (Very Strong Penalization), the following penalized criteria are assessed:

- AIC like: $\frac{1}{n}\sum_{t=1}^{n}\rho_{0.5}\left(z_t - F_\theta(x_t, y_t)\right) \times \left(1 + \frac{2D}{n}\right)$
- BIC like: $\frac{1}{n}\sum_{t=1}^{n}\rho_{0.5}\left(z_t - F_\theta(x_t, y_t)\right) \times \left(1 + \frac{D\log n}{n}\right)$
- SP (Strong Penalization): $\frac{1}{n}\sum_{t=1}^{n}\rho_{0.5}\left(z_t - F_\theta(x_t, y_t)\right) \times \left(1 + \frac{D\sqrt{n}}{n}\right)$

We simulate $n = 100$, $n = 500$ and $n = 1000$ data according to the true model (9), for each n the experiment is repeated 100 times.

The following architectures are chosen by the penalized criteria :

- n=100

	nb h. units	1	2	3	4	5	6	7	8	9	10
AIC like	models sel.	0	0	**13**	**10**	**5**	**6**	**2**	**10**	**21**	**33**
BIC like	models sel.	0	**9**	**86**	**3**	0	**1**	0	0	0	**1**
SP	models sel.	**3**	**36**	**61**	0	0	0	0	0	0	0

- n=500

	nb h. units	1	2	3	4	5	6	7	8	9	10
AIC like	models sel.	0	0	**62**	**19**	**13**	**5**	**1**	0	0	0
BIC like	models sel.	0	0	**100**	0	0	0	0	0	0	0
SP	models sel.	0	**2**	**98**	0	0	0	0	0	0	0

- n=1000

	nb h. units	1	2	3	4	5	6	7	8	9	10
AIC like	models sel.	0	0	**72**	**13**	**7**	**6**	0	**2**	0	0
BIC like	models sel.	0	0	**100**	0	0	0	0	0	0	0
SP	models sel.	0	0	**100**	0	0	0	0	0	0	0

The BIC like criterion and the Strong Penalization chose often the true architecture even for a small number of data. According to the theory, AIC like criterion is not consistent (see condition **A1**) and the chosen architecture is always too large. The Strong penalization chose a too small architecture when the number of data is small ($n = 100$), however it is a consistent criterion, so its behavior is correct for larger number of data ($n = 500$ and $n = 1000$). The BIC like criterion seems to be the best for this cost function.

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0.
Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

4 Conclusion

The conventional least squares estimator may be seriously deficient in case of non-Gaussian errors. It seems reasonable to pay a small premium in the form of sacrificed efficiency, in order to get more robust regression models. The class of statistics model called "regression quantiles" are known to have good properties under some restrictive assumptions. In this paper we have shown that some results may be obtained under more general assumptions. We have proven an inequality showing that overfitting of theses models is moderate if the noise admits exponential moments. This bound justifies the use of penalized criterion similar to the BIC criterion in order to fit the dimension of models. Finally, a more challenging task may be to get a more precise tuning of penalization term which, according to our result, can be chosen among a wide range of functions.

References

[1] Koenker, R. and Basset, G., Regression quantiles. *Econometrica*, 46:1, pages 33-50, 1978

[2] Engel, E., Die produktions- und Konsumptionverhaltnisse des Konigreichs Sachsen.*International Statistical Institute Bulletin*, 9, pages 1-125, 1857

[3] J. Rynkiewicz, Consistent estimation of the architecture of multilayer perceptrons. In M. Verleysen, editor, *proceedings of the* 14^{th} *European Symposium on Artificial Neural Networks* (ESANN 2006), d-side pub., pages 149-154, April 28-30, Bruges (Belgium), 2006.

[4] J. Rynkiewicz, General bound of overfitting for MLP regression models. *Neurocomputing* to appear.

[5] A.W. van der Vaart, *Asymptotic statistics*, Cambridge university Press, Cambridge, 1998.

[6] Weiss, A., Estimating nonlinear dynamic models using least absolute error estimation.*Econometric Theory*, 7, pages 46-68, 1991 *Econometrica*, 46:1, pages 33-50, 1978