ON KERNEL-BASED NONLINEAR REGRESSION ESTIMATION

Jan Kalina – Petra Vidnerová

Abstract

This paper is devoted to two important kernel-based tools of nonlinear regression: the Nadaraya-Watson estimator, which can be characterized as a successful statistical method in various econometric applications, and regularization networks, which represent machine learning tools very rarely used in econometric modeling. This paper recalls both approaches and describes their common features as well as differences. For the Nadaraya-Watson estimator, we explain its connection to the conditional expectation of the response variable. Our main contribution is numerical analysis of suitable data with an economic motivation and a comparison of the two nonlinear regression tools. Our computations reveal some tools for the Nadaraya-Watson in R software to be unreliable, others not prepared for a routine usage. On the other hand, the regression modeling by means of regularization networks is much simpler and also turns out to be more reliable in our examples. These also bring unique evidence revealing the need for a careful choice of the parameters of regularization networks.

Key words: Nonlinear regression, machine learning, kernel smoothing, regularization, regularization networks

JEL Code: C20, C45, C63

Introduction

Regression modeling is well known to represent a fundamental task in econometrics. While the importance of its nonlinear versions has been acknowledged already decades ago (Fisher & Salmon, 1986), practical applications of nonlinear models in econometrics have become popular only rather recently (Racine, 2019). This work is devoted to two important methods of nonlinear regression with an unknown regression function. These include the Nadaraya-Watson (shortly N-W) estimator and regularization networks.

The N-W estimator has found numerous econometric applications and we now recall some of the recent ones. For example, Kibria et al. (2019) applied the N-W estimator in a study of different sources of electrical energy in various countries and focused on the

relationship between the share of fossil fuels and the economic growth. The N-W estimator was used for option pricing in Kenmoe & Sanfelici (2014), or in the study of audit quality effect by Yang et al. (2020). Other available applications include an extension of the N-W estimator for functional data was used within the study of the aggregate stock market by Hong & Linton (2020). An adaptive version of the N-W estimator with a varying size (i.e. varying bandwidth) of the kernel was used e.g. by Ahmed et al. (2020) for estimating nonlinear regression quantiles.

Regularization is commonly used in machine learning in order to solve ill-posed problems in various tasks of regression, classification, clustering, or dimensionality reduction (Kalina & Schlenker, 2015). This paper studies regularization networks, which represent a particular type of machine learning tools also denoted as regularized ridge estimators. To avoid misunderstanding, we do not discuss regularized networks (i.e. regularized versions of common types of networks) here. We are not aware of economic applications of regularization networks; applications or regularization networks to the context of dynamical systems (however without a direct economic example) were presented by Chiuso & Pillonett (2019). There, the method was used for learning from examples using a nonlinear black-box system, while the complexity of the models was controled by the regularization parameter.

The nonlinear regression model assumes the total number n of observations to be available. We denote the values of the continuous response variables as $Y_1, \dots, Y_n$ and corresponding values of p-variate regressors (independent variables) as $X_1, \dots, X_n$, where $X_i = (X_{i1}, \dots, X_{ip})^T$ for $i = 1, \dots, n$. The aim is to find the best approximation of the response conditioning on the regressors. Particularly, we consider the nonlinear regression model

$$Y_i = f\left(\beta_1 X_{i1}, \dots, \beta_p X_{ip}\right) + e_i, \quad i = 1, \dots n, \tag{1}$$

with an unknown function f, where $e_1, \dots, e_n$ are random errors in the model and $\beta = (\beta_1, \dots, \beta_p)^T$ is a vector of parameters.

We can say that both the N-W estimator and regularization networks belong to nonparametric approaches, as they do assume specific distribution of the errors in (1) and as the predicted value of the response is an unknown parameter for each value of the regressor. In the regression framework, the task of nonparametric regression can be described as function estimation (function approximation), and its aim is to estimate a continuous function of the regressors, while the value of the function for each possible value of the regressors is an unknown parameter (Kalina, 2014).

1 Nadaraya-Watson estimator

The N-W estimator represents a kernel-based regression method popular in the regression setup (1), but also in models without any regressors; in the latter case, the estimator is known as the kernel smoother. This section recalls the estimator and then desribes a related kernel-based estimator of conditional expectation in Section 1.1, which is straightforward but still difficult to be found in the literature.

Known properties of the N-W estimators include their consistency, asymptotic bias, or asymptotic normality. These results have been derived only under rather stringent assumptions, just like theoretical comparisons of various kernels. In practice, it is however not possible to verify if the assumptions are fulfilled. Usually, one assumes the observations to correspond to realizations of independent identically random vectors. Let us stress here that regressors are assumed to be random. Further, it is common to assume, apart from other technical assumptions, that $EY^2 < \infty$, where Y denotes the random response variable.

Computing the N-W estimator requires to specify a kernel K, which serves to measure the distance in the p-dimensional space; common choices include the Gaussian kernel or Epanechnikov kernel, which itself exists in several different versions. At any case, the N-W estimator uses a normalized version of the kernel obtained as

$$\widetilde{K}(x) = h^{-p} K\left(\frac{x}{h}\right), \quad x \in \mathbb{R}^p, \tag{2}$$

where the parameter h (bandwidth) fulfilling $h > 0$ corresponds to the level of smoothing. The N-W estimator of the function f in (1) with a normalized kernel (2) has the form

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} Y_i \widetilde{K}(x - X_i)}{\sum_{i=1}^{n} \widetilde{K}(x - X_i)}, \qquad x \in \mathbb{R}^p, \tag{3}$$

i.e. represents a linear function of the response. Thus, it commonly denoted as a linear estimator, although it is not a linear estimator of f ($\hat{f}$ is not a linear function of $x \in \mathbb{R}^p$).

From the practical point of view, the influence of the choice of K turns out not to be so strong, so it remains crucial to select a proper h. There are no explicit expressions for its optimal value. Formally, numerous criteria for the optimal h have been derived, which however depend on unknown parameters; for practical purposes, cross validation may be recommended to find a suitable (but not the optimal) h. The N-W estimator requires a large number n of observations; especially if the dimensionality p increases, the number n required for retaining the same approximation ability increases exponentially. Such phenomenon is known as the curse of dimensionality.

1.1 A kernel-based estimator of the conditional expectation

Let us now show the connection of the N-W estimator with the conditional expectation. We understand here the regressors as realizations of a *p*-dimensional random vector, which is denoted here as *X*. We may formulate an optimization task

$$m \coloneqq \arg\min_{f \in M} E\big(Y - f(X)\big)^2; \tag{4}$$

here, M denotes the class of real measurable functions, where each function $f \in M$ fulfils $Ef^2 < \infty$. Let us stress that we consider expectations under the assumption that the distribution of the random variables is known.

The solution of the optimization task (4) is precisely the conditional expectation in the form $m(x) = E(Y|X = x)$, where $x \in \mathbb{R}^p$. In practice, however, the conditional expectation is not usually used as an estimate of f, as its computation can be very tedious. The task (4) may be replaced by a close (approximated) task exploiting the relation

$$m(x) = E(Y|X = x) = \int y f(y|x) dy = \int y \, \frac{f(x,y)}{f_x(x)} dy, \tag{5}$$

where $f(x, y)$ denotes the joint density and f_x the marginal density of the regressors. Such approximation of the function *m* has a form

$$\widehat{m}(x) = \sum_{i=1}^{n} Y_i \frac{\sum_{j=1}^{n} \widetilde{K}_y(y - Y_j)}{\sum_{j=1}^{n} \widetilde{K}(x - X_j)}, \quad \text{where} \quad \widetilde{K}_y(y) = \frac{1}{h} K\left(\frac{y}{h}\right); \tag{6}$$

this requires to use kernel estimates of densities for the numerator and denominator as

$$\hat{f}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \widetilde{K}(x - X_i)\, \widetilde{K}_y(y - Y_i) \quad \text{and} \quad \hat{f}_x(x) = \frac{1}{n} \sum_{i=1}^{n} \widetilde{K}(x - X_i) \tag{7}$$

for $x \in \mathbb{R}^p$ and $y \in \mathbb{R}$, respectively. It is already evident that the obtained nonlinear estimate (6) of the conditional expectation is related to the N-W estimator (3).

2 Regularization networks

Regularization networks proposed by Girosi et al. (1995) may be characterized as nonlinear regression tools with a clear interpretation and a straightforward computation; in fact, they may be computed using an available explicit formula. Still, as stated already in the Introduction, regularization networks are currently avoided in econometric applications. This is perhaps because of their simplicity, or also because their theoretical properties have not

been known. Actually, the performance of regularization networks has been studied mainly by means of numerical experiments (Neruda & Vidnerová, 2009).

Assuming again the nonlinear regression model (1), let us now formulate the task of finding an unknown continuous function *f* as argument of minima

$$\min_f \sum_{i=1}^n (Y_i - f(X_i))^2, \tag{8}$$

which is considered over the space of all continuous functions. This task of nonparametric regression is ill-posed, because there exist more than one solution. In fact, there exist an infinite number of solutions and one of them is the trivial (degenerated) solution in the form of a piece-wise linear function connecting all values of the response. The following approach assumes that the user chooses some kernel; the most common choice is the Gaussian kernel (2). It follows from the theory of functional analysis that a space (say H_K) of functions corresponds to every fixed *K*.

When a particular kernel is chosen, we can replace the original task (8) by a new task

$$\min_{f \in H_K} \sum_{i=1}^n (Y_i - f(X_i))^2, \tag{9}$$

which still remains to be ill-posed. Thus, we consider yet another task

$$\min_{f \in H_K} \left\{ \sum_{i=1}^n (Y_i - f(X_i))^2 + \lambda ||f||_{H_K} \right\}, \quad \lambda > 0, \tag{10}$$

where

$$||f||_{H_K} = \sum_{i=1}^n \sum_{j=1}^n \beta_i \beta_j K(X_i, X_j). \tag{11}$$

Such metric in the functional space evaluated the distance between *p*-variate vectors by means of the kernel *K*.

We skip theoretical results from the field of functional analysis, devoted to the space of real functions or RKHS (Reproducible Kernel Hilbert Space) spaces (Neruda & Vidnerová, 2009). We will only need here the representation theorem (Girosi et al., 1995) stating that the resulting estimator can be expressed as

$$\hat{f}(x) = \sum_{i=1}^n \beta_i K(x, X_i), \quad x \in \mathbb{R}^p, \tag{12}$$

depending on the vector of parameters $\beta = (\beta_1, \dots, \beta_n)^T$. It will be convenient to denote by *K* the square matrix $K = \left(K(X_i, X_j)\right)_{i,j=1}^n$. Then, the task (10) can be expressed as

$$\min_{\beta}\{||Y - K\beta||^2 + \lambda\beta^T K\beta\}. \quad (13)$$

By using derivatives, we find out already easily that the explicit form of the minimum is

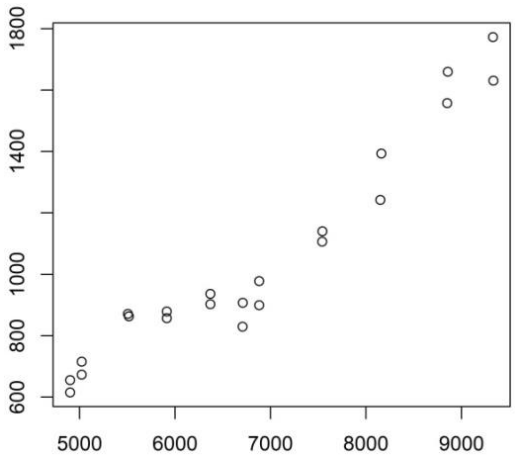
$$\hat{\beta} = (K^T K + \lambda K)^{-1} K^T Y = [(K + \lambda I)K]^{-1} K^T Y = (K + \lambda I)^{-1} Y. \quad (14)$$

The estimate $\hat{\beta}$ is known as a regularization network, although statisticians often prefer to call it a generalized ridge estimator (Mori & Suzuki, 2018); the latter reminds the connection between (14) and the ridge regression estimator (in linear regression). It is necessary to recall that the properties depend not only on λ, but also on the variance of the Gaussian kernel (say σ^2). Finally, the fitted value for a particular $x \in \mathbb{R}^p$ is obtained as an empirical counterpart of (12) in the form $\hat{f}(x) = \sum_{j=1}^{n} \hat{\beta}_j K(x, X_j)$.

3 Examples

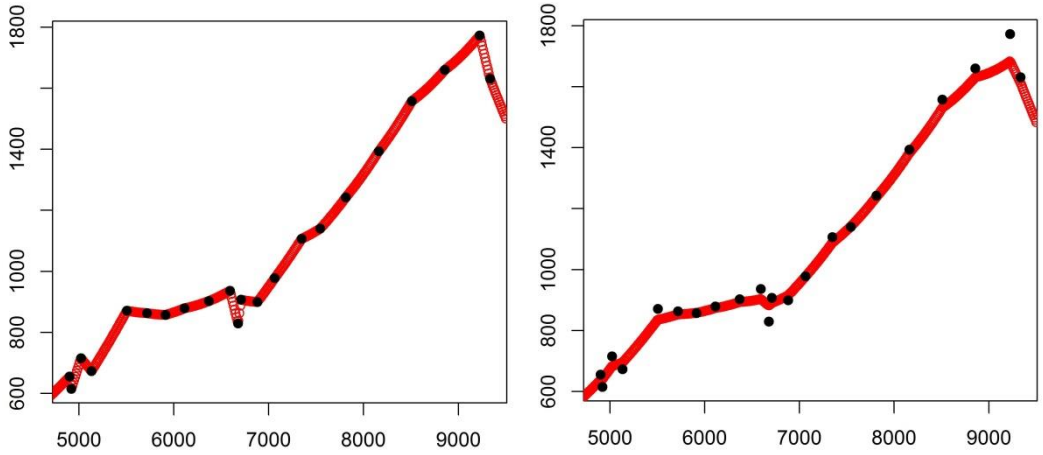
This section presents our computations of the N-W estimator and regularization networks over two datasets denoted here as dataset A and dataset B. In fact, we are not aware of comparisons of these two methods over real economic data. In the dataset A with $n = 23$ and $p = 1$, the response variable corresponds to real gross private domestic investments in the United States and a single regressor is the GDP, while both variables are yearly measurements from the years 1980-2001 in 10^9 USD. We create here an artificial dataset B with $n = 22$ and again $p = 1$, which is shown in Figure 1.

Fig. 1: Dataset B. Horizontal axis: the regressor (GDP). Vertical axis: the response (investments).



Source: own artificial data.

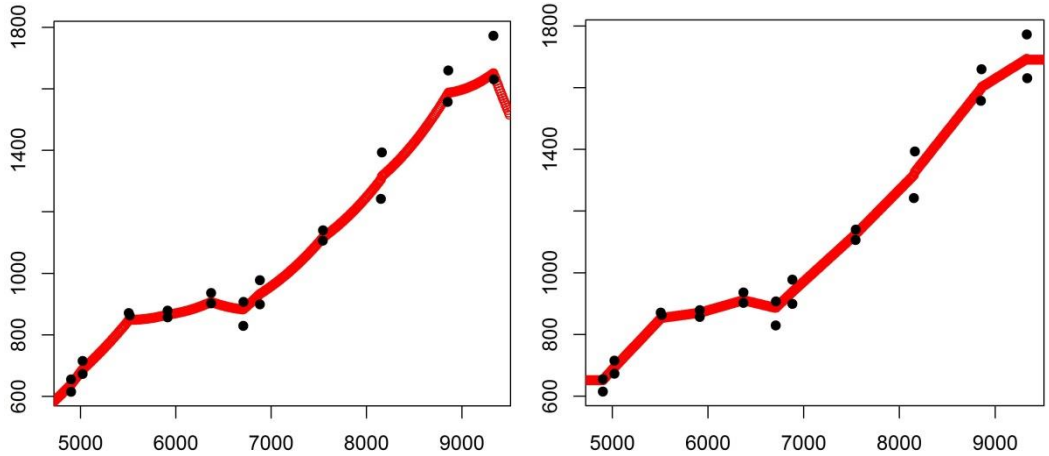
Fig. 2: Regularization networks for dataset A with σ^2=1000. Horizontal axis: GDP. Vertical axis: investments. Left: $\lambda = 0$. Right: $\lambda = 0.1$.



Source: own computation.

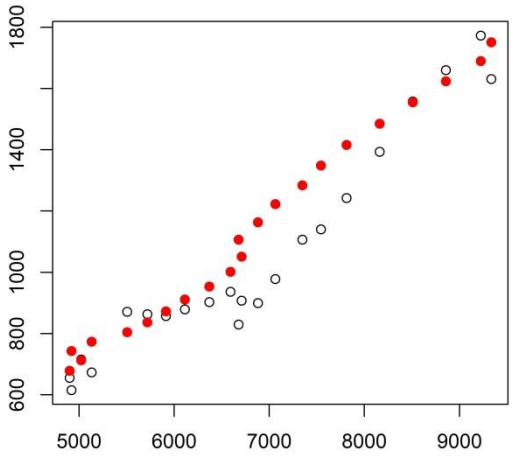
Results of regularization networks with different values of their parameters are shown in Figures 2 and 3. The fit with $\lambda = 0$ in Figure 2 (left) does not use any regularization (i.e. is extremely overfitted); thanks to the effect of a positive λ, the fit in Figure 2 (right) is much smoother. The two images of Figure 3 are shown as a warning; the left image yields a suitable fit with reasonable values of the parameters. The right image however has extremely selected parameters, which themselves are definitely unsuitable for the data, but the fit still seems very good (in fact smoother than in the left image). Thus, the user should be very careful and we do not recommend to rely on automatic procedures for computing regularization networks without interpreting their parameters.

Fig. 3: Regularization networks for dataset B. Left: $\lambda = 0.1$ and σ^2=1000. Right: $\lambda = 0.0001$ and σ^2=10^7. Horizontal axis: GDP. Vertical axis: investments.



Source: own computation.

Fig. 4: The Nadaraya-Watson estimator (full circles) computed for dataset A; raw data shown are as empty circles. Horizontal axis: GDP. Vertical axis: investments.



Source: own computation.

We used three methods for computing the N-W estimator, using three different libraries of R software. Library bbemkr (function NadarayaWatsonkernel) yields no result, presumably because of a small *n*; library stats (function ksmooth) yields no result again; and finally using library monreg (function monreg), the fitted values of the response are shown in Figure 4. We consider the results obtained by the library monreg to be quite poor. The result was obtained with the Epanechnikov kernel with a default value of the bandwidth. This library, just like the libraries bbemkr and stats, do not contain tools for a cross-validated search of a suitable bandwidth. Thus, we do not consider these tools to be user-friendly, as the user would has to perform the search for a suitable bandwidth only manually. In addition, it should be mentioned that a user-friendly method for computing the N-W estimator together with a cross validation for finding the optimal bandwidth is available in the Python programming language, namely in the package mcmodels.

Finally, let us conclude that the regularization networks clearly estimate the trend much better compared to the N-W estimator. Although some of the images seemingly reveal the obtained trend to correspond to the data very well, such autovalidation (autoverification) only reveals here to fit the training data well, while the images do not show if the method is able to generalize well to new data or not.

Conclusions

This paper starts with a detailed comparative description of the N-W estimator and regularization networks. Both are flexible nonlinear regression tools; at the same time, their

nonparametric structure ensures robustness (as in Saleh et al. (2012)). The N-W estimator has been applied to many economic data analysis tasks, but its performance in our numerical examples is outperformed by (much less known) regularization networks. In addition, we realize here even on a (simplistic) economic dataset that the methods heavily depend on the choice of the parameters. Thus, both methods require a careful modeling over real data; we stress here that their parameters (commonly denoted as hyperparameters) should be checked to be meaningful for the data under consideration. This complicates their potential implementation within automatic procedures; we are actually not aware of a publicly available software procedure for their fully automatic computation. While there exist numerous other nonlinear regression methods within statistics as well as machine learning, it turns out that their systematic comparisons over economic data, including a minimization of the cross validation mean square error, would be very useful.

Acknowledgment

The work was supported by the project GA21-05325S ("Modern nonparametric methods in econometrics") of the Czech Science Foundation.

References

Ahmed, H.I.E.S., Salha, R.B., & El-Sayed, H.O. (2020). Adaptive weighted Nadaraya-Watson estimation of the conditional quantiles by varying bandwidth. *Communications in Statistics-Simulation and Computation*, 49, 1105-1117.

Chiuso, A. & Pillonetto, G. (2019). System identification: A machine learning perspective. *Annual Review of Control, Robotics, and Autonomous Systems*, 2, 281-304.

Fisher, P. & Salmon, M. (1986). On evaluating the importance of nonlinearity in large macroeconometric models. *International Economic Review*, 27, 625-646.

Girosi, F., Jones, M., & Poggio, T. (1995). Regularization theory and neural networks architectures. *Neural Computation*, 7, 219-269.

Hong, S.Y. & Linton, O. (2020). Nonparametric estimation of infinite order regression and its application to the risk-return tradeoff. *Journal of Econometrics*, 219, 389-424.

Kalina, J. (2013). Highly robust methods in data mining. *Serbian Journal of Management*, 8, 9-24.

Kalina, J. (2014). On robust information extraction from high-dimensional data. *Serbian Journal of Management*, 9, 131-144.

Kalina, J. & Schlenker, A. (2015). A robust supervised variable selection for noisy high-dimensional data. *BioMed Research International*, 2015, Article 320385.

Kenmoe, R.N. & Sanfelici, S. (2014). An application of nonparametric volatility estimators to option pricing. *Decisions in Economics and Finance*, 37, 393-412.

Kibria, A., Akhundjanov, S.B., & Oladi, R. (2019). Fossil fuel share in the energy mix and economic growth. *International Review of Economics and Finance*, 59, 253-264.

Mori, Y. & Suzuki, T. (2018). Generalized ridge estimator and model selection criteria in multivariate linear regression. *Journal of Multivariate Analysis*, 165, 243-261.

Neruda, R. & Vidnerová, P. (2009). Learning errors by radial basis function neural networks and regularization networks. *International Journal of Grid and Distributed Computing*, 1, 49-57.

Racine, J.S. (2019). *An introduction to the advanced theory and practice of nonparametric econometrics: A replicable approach using R*. Cambridge: Cambridge University Press.

Saleh A., Picek J., & Kalina J. (2012). R-estimation of the parameters of a multiple regression model with measurement errors. *Metrika*, 75, 311-328.

Yang, J.C., Chuang, H.C., & Kuan, C.M. (2020). Double machine learning with gradient boosting and its application to the Big N audit quality effect. *Journal of Econometrics*, 216, 268-283.

Contact

Jan Kalina
The Czech Academy of Sciences, Institute of Information Theory and Automation
Pod Vodárenskou věží 4, 182 00 Prague 8, Czech Republic
& The Czech Academy of Sciences, Institute of Computer Science
Pod Vodárenskou věží 2, 182 07 Prague 8, Czech Republic
kalina@cs.cas.cz

Petra Vidnerová
The Czech Academy of Sciences, Institute of Computer Science
Pod Vodárenskou věží 2, 182 07, Praha 8, Czech Republic
petra@cs.cas.cz