N-KERNEL: A REVIEW

MIGUEL A. DELGADO
Department of Economics, Indiana University, Bloomington; IN 47405, U.S.A

AND

THANASIS STENGOS
Department of Economics, Guelph University, Guelph, Ontario N1G 2W1, Canada

SUMMARY

The package performs estimation and prediction in the context of time-series or cross-section nonparametric models. It is menu-driven and very easy to operate. The manual reads well. This version has some limitations, which can easily be corrected. Nevertheless it provides a useful pedagogical and research tool, even for people not familiar with nonparametric analysis.

1. INTRODUCTION

There is considerable current interest in nonparametric estimation and testing of econometric functionals. Robinson (1986) discusses how nonparametric estimates of densities and conditional expectations may be used in model specification. Nonparametric estimates are also used in the estimation of semiparametric models where it is explicitly recognized that certain features of the underlying distribution of the data are unknown while other features follow a known parametric model. The goal in this case is to obtain estimates for the parametric part asymptotically equivalent to those obtained when the nonparametric part of the model is known (see Robinson, 1987, for a survey).

The aim of N-Kernel, hereafter NK, is to provide a menu-driven package to perform nonparametric regression analysis, either with a cross-sectional data set or within a time-series framework. The estimation method is due to MacQueen (1979).

The program, as its name suggests, uses kernel density estimation as the main building block on which estimation of more complicated functionals, including regression functionals, is based. Intuitively, NK density estimation fits a smooth density to a multivariate sample of N observations and D variables, one of which is taken to be the regressand and all or some of the remaining $D-1$ variables are the regressors. The NK density is a mixture of N multivariate normal densities, one for each sample point. Each component normal has its own mean and covariance structure. These are determined by the sample points in the neighbourhood of the sample point in question. This neighbourhood of points adjacent to each sample point is controlled through the parameter GR, the 'group radius'. This parameter controls the smoothness of the density estimate. Large values of GR lead to a single multivariate normal density being fitted to the whole sample, whereas smaller values of GR lead to a mixture of multivariate normals, where each component of this mixture is fitted to merely the points allowed by GR. Note that the mixture density is quite smooth. Because of the averaging that

takes place, its derivative tends to be smaller than the derivatives of the individual components that tend to fluctuate more.

2. THE ESTIMATION METHOD

Formally, given a sample $\{X_1, \ldots, X_n\}$ from the R^D random variable X, the density estimate evaluated at z is estimated by,

$$\hat{f}(z) = n^{-1} \sum_i |A_i|^{-1/2} \phi(A_i^{-1/2}(z - Z_i)) \tag{1}$$

where the summation runs from 1 to n, $\phi(\cdot)$ is the multivariate standard normal density, and

$$A_i = \left(\sum_j X_j X_j' \omega_{ij} - \sum_j X_j \omega_{ij} \sum_j X_j' \omega_{ij} \right) b^2 \tag{2}$$

$$Z_i = b \sum_j X_j \omega_{ij} + (1 - b)\bar{X} \tag{3}$$

where $\bar{X} = n^{-1} \Sigma_i X_i$, $b = \{n/(n-1)\}^{1/2}$ and

$$\omega_{ij} = 1(i \neq j) C_{ij} C^{-1} + 1(i = j) \left\{ 1 + \left(1 - \sum_j C_{ij} \right) C^{-1} \right\} \tag{4}$$

where $1(\cdot)$ is the usual indicator function, $C = \max_i \Sigma_j C_{ij}$ and $C_{ij} = 1(\rho(X_i, X_j) \leqslant \mathrm{GR})$, where GR is the 'group radius', chosen by the user, and $\rho(X_i, X_j) = \max_{1 \leqslant k \leqslant D} |X_{ik} - X_{jk}|/s_k$, where s_k is the sample standard deviation of the kth X component.

Note that $\Sigma_j \omega_{ij} = \Sigma_i \omega_{ij} = 1$ and, therefore, the random variable with density $\hat{f}(\cdot)$ has mean $\bar{X}$ and variance equal to the sample covariance of X. This is why the density in (1) produces unbiased statistics.

A feature of the estimation method used in the NK program which distinguishes it from other more standard kernel-based methods is that NK uses multivariate kernels that fluctuate from point to point with a variable covariance structure at each sample point. More traditional kernel methods use a product of normals with a fixed standard deviation in each dimension in order to construct this mixture density; i.e. the density at point z is estimated by

$$\tilde{f}(z) = (\sigma^D n)^{-1} \sum_i \phi\{(z - X_i)/\sigma\} \tag{5}$$

That is, the density estimate in this case is a mixture of multivariate normals with different means and a diagonal covariance matrix with equal components.

An advantage of the variable kernel approach adopted by NK is that it avoids the sometimes spurious bumps that one observes in the tails of the empirical distribution with the product kernel. The latter, of course, requires much less computation.

In applying NK to the regression problem, an estimate $E_N(Y \mid X = z)$ is estimated from the estimated joint distribution of Y and X simply by conditioning on $X = z$. Letting L_i stand for the (linear) regression function of Y given X in the ith normal component of estimated distribution, then

$$E_N(Y \mid X = z) = \hat{f}(z)^{-1} n^{-1} \sum_i L_i |A_i|^{-1/2} \phi(A_i^{-1/2}(z - Z_i)) \tag{6}$$

This regression function behaves differently from the regression computed from the product

kernel. The author suggests that it may be an improvement over the latter, because the 'local' kernels are sensitive to the variation in the regression surface over different subregions of the sample. The average slopes of the component regression lines are printed out in order to provide a basis for assessing the relative contribution of the variables to the nonlinear regression.

The choice of the parameter GR is, of course, one of the central issues in the NK regression analysis. A choice of large GR leads to over-smoothing and the introduction of bias, whereas too small a GR might lead to imprecise results due to the extra noise. The author suggests a trial-and-error method for choosing GR, observing a nonsingularity constraint such that the neighbourhood of points around a given sample point contains enough points to define a non-degenerate multivariate normal density. However, one has to be aware that the regression results that one obtains are conditional on the choice of GR, and at times the results might change quite a lot with different GR choices. There is, however, a nice pedagogical feature with regard to the choice of the GR that can be of value to the applied researcher. A large enough value of GR leads to a single multivariate normal density being fitted on the whole sample. This is the case of a linear regression function, assuming joint normality of the dependent and independent variables. Hence, one can easily compare the NK results from smaller GRs to the one that corresponds to the linear regression case. One can then assess the extent to which the NK estimator is able to capture the underlying nonlinear regression relationship. Hence, one of the important features of NK is to provide a benchmark to check the performance of a linear model.

However, one has to be careful to interpret the genuine nonlinearity that the NK estimator captures. This is done by means of estimating the NK-regression function from an artificially created 'randomized sample', where any possibly genuine association between the dependent variable and the independent ones is destroyed, and then comparing it to the NK-regression function estimate from the original sample. Any structure that is captured in the randomized sample is spurious and is the result of 'overfitting'. What is genuine is the difference in what the NK estimator captures in the original sample and what it captures in the randomized samples. An additional evaluation method is provided by 'cross-validation', where one uses the estimated regression function from a subsample to predict the actual values of the dependent variable in another subsample.

3. SOFTWARE CAPABILITIES

The present version of NK is designed to work with 640K of RAM with a maximum number of 620 observations and a maximum number of 24 variables allowed. The 'randomized sample' and 'cross-validation' procedures are performed with the aid of auxiliary programs that are compiled separately in order to avoid using up any RAM when NK is loaded. The price of this separation is that one has to leave the main environment in order to use one of these auxiliary programs. The main drawback with the 640K RAM limitation is the constraint it imposes on the allowable size of the data set. Since the use of NK is to provide a guidance for the detection of possible nonlinearities, the latter are likely to arise in typically large cross-sections or time-series data of higher frequency such as daily or weekly data. Hence, if NK is to be anything beyond a pedagogical tool for applied econometric research the above constraint will have to be relaxed. Given recent advances in microprocessor technology the extension of the program to work with 1M or more should not pose too much of problem.

The main menu options are: input the data points; edit or append the data matrix; perform transformations on data vectors; examine data matrix (descriptive statistics); perform

regression analysis; perform time-series analysis. Each of these options offers different menus which are self-explanatory and are well discussed in the manual.

The data input can be a ASCII file with at least two columns, each of the columns being observations of two variables, or it may be introduced directly from the terminal. An existing file may be edited in order to introduce new observations, deleting existing ones, etc.

The data manipulation capabilities of the program can be improved upon. It only permits a couple of transformations of the series, logs and product of a series by another.

The data display options are adequate. All or part of the data matrix, as well as summary statistics, can be printed. Scatterplots of the data, including lagged values are offered. With a EGA or VGA card the program produces good full-colour graphs. The graphs in Figure 1 (parts a, b and c) have been produced using a 24-pitch printer.

When the regression option is chosen, the program asks for the name or position of the dependent and independent variables in the data matrix as well as the value of GR.

One of the good features of the package is the analysis of the results. There is a concerted effort to provide a lot of 'summary statistics' of the results of the regression analysis in more or less the same way as one would obtain them from a standard linear regression package. There are measures of fit such as R^2, and a detailed analysis of the summary statistics of the residuals, including mean, variance and root mean squared error. The residuals can be stored for further analysis. These statistics provide good first-hand impression of the performance of the NK estimator of the regression function. However, one should bear in mind that these statistics are conditional on the choice of the GR parameter. In order to judge the overall performance of the NK estimates one has to take into account the effect of over-fitting by using the 'randomized sample' evaluation procedure discussed earlier. On the negative side, in presenting the results there is no proper accounting for the derivatives of the regression function with respect to the variables, or as we know them the 'beta' coefficients in the case of the linear model. It is precisely these coefficients that the researcher cares about, and although they vary from sample point to sample point it would be helpful to have them not only as an average but also evaluated at specific points of the sample space of the independent variable in question, such as its sample mean, quantiles or some other point of its distribution. It is a deficiency of the program not to provide proper standard errors to assess the significance of these estimated derivatives (estimated betas), even though it is to be expected that these estimates are going to be imprecise.

Besides the analysis of residuals, the regression options include the calculation of predicted values with the independent variables values input from the terminal or from a data file, plots of regression curves and conditional probability density and calculation of conditional probabilities. The GR value can be changed, if desired, within the regression menu.

The graphics capabilities of the regression menu are adequate. The package produces graphs of the conditional expectation estimates with their corresponding confidence bounds. The confidence bounds are computed from nonparametric estimates of the conditional standard deviations based on NK nonparametric estimates of the first and second conditional moments of the dependent variable. In Figure 1 (parts a, b and c) we report a hard copy of the conditional estimates plots using a sample data (S2A30) included in the package. The true conditional expectation follows a logistic curve. These data consist of two variables with 30 observations. We have just regressed one variable against other using 3 GR values, GR = 1, 2 and 5. Note that as GR increases the conditional expectation estimates become more smooth. When GR = 5, $C_{ij} = 1$ for all i, j and the fitted regression line corresponds to the linear regression model estimated by ordinary least-squares. Plots of the conditional densities are also offered. However, there is no provision of plots of the unconditional densities. Frequently,

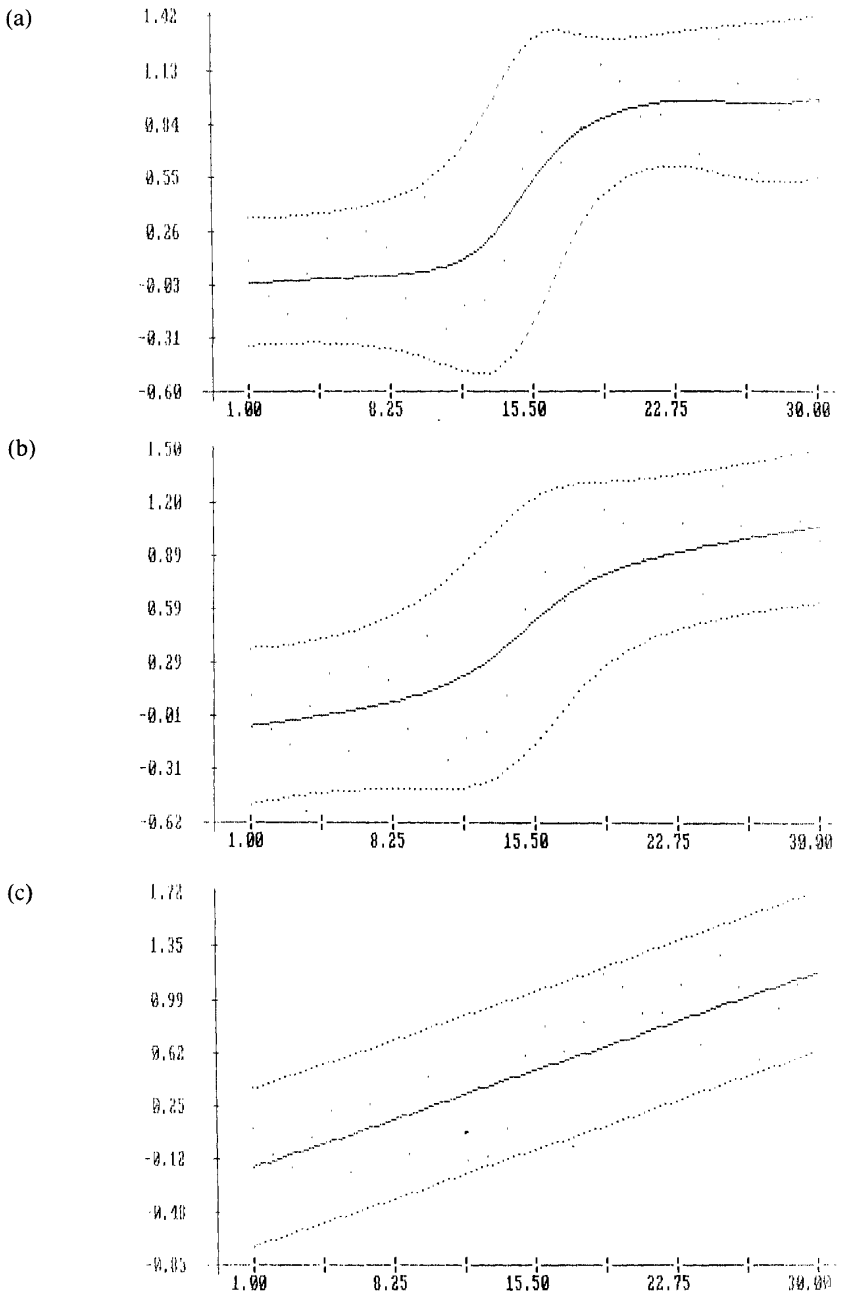

Figure 1. NK conditional expectation estimates of $E[Y|X] = 1/(1 + \exp(15 \cdot 5 - X))$. (a) GR = 1; (b) GR = 2; (c) GR = 5

these plots are as useful and informative as conditional plots. Besides, a very informative way to summarize the results of the analysis would be to use plots of the unconditional densities of the residuals and/or the empirical distribution of the betas. In our opinion this is an oversight that should be corrected in future versions of the program.

The time-series menu permits one to define as regressors lags of the dependent or independent variables, thereby allowing estimation of distributed lag models. Besides the usual output, NK offers forecast values for any specified number of periods.

4. CONCLUSIONS

The program seems especially adequate for investigating the performance of a given parametric regression model. The conditional density plots are also helpful for checking departures from normality of the data.

The package would become more valuable if it were to report additional nonparametric estimates. In particular, unconditional density plots are helpful. Some users may also be interested in traditional kernel estimates of the conditional expectations and/or densities. For instance the statistical properties of semiparametric estimates have been derived using density estimates as (5) or nearest neighbours probabilistic weights. Therefore estimates based on (5) may be useful in semiparametric analysis. It does not seem very much of a problem to introduce these extra estimates.

The disk, together with the manual, are available from: Non-standard Statistical Software, 513 Wilshire Blvd., Suite 311, Santa Monica, CA 90401, USA, at a price of $49 plus $3 postal and handling charges. There is a HELP LINE consisting of a telephone with a tape-recorder available at: 213 450-9313.

REFERENCES

MacQueen, J. B. (1979). 'On kernel random variables with unbiased statistics', Western Management Science Institute Working Paper No. 286.

Robinson, P. M. (1986). 'Nonparametric methods in specification', *Economic Journal*, Suppl. 96, pp. 134–141.

Robinson, P. M. (1987). 'Semiparametric econometrics: a survey', *Journal of Applied Econometrics*, **3**, 35–51.