NONPARAMETRIC ESTIMATION

Maarten Jansen [1], Université libre de Bruxelles, Belgium
Gerda Claeskens [2] , KU Leuven, Belgium

1 Meanings of the word "nonparametric"

The terminology *nonparametric* has been introduced by Wolfowitz in 1942 to encompass a group of statistical techniques for situations where one does not specify the functional form of the distributions of the random variables that one is dealing with. In its earlier form, this comprised mainly methods working with rank statistics, and was also coined "*distribution free*" methods. Most often these methods are applied to perform hypothesis tests. For an example of such a hypothesis test, see the entry by Jurečková (same volume). Other examples include the Kolmogorov-Smirnov test, the runs test, sign test, Wilcoxon signed rank test, the Mann-Whitney U-test and Fisher's exact test. For an overview and details, see Hollander and Wolfowitz (1999). This type of nonparametric methods has the advantage that it can be applied to ordinal and rank data; the data may be frequencies or counts, and do not have to be measured on a continuous scale.

In more recent times, nonparametric statistics has evolved to settings where a model for the data is not specified a priori, but is in some form determined from the data. This will be explained in more detail below. In such nonparametric models there are parameters to estimate, even many parameters in most cases, hence the name-giving might be somewhat misleading. Main examples of such estimation methods are kernel estimators, splines and wavelet estimators. These techniques are also known as "*smoothing methods*".

2 Nonparametric regression estimation

In parametric regression models we relate the mean of a response variable Y, conditional on covariates X via a parametric function. For example, in linear regression models, we assume that $Y = \beta_0 + X\beta_1 + \varepsilon$, where β_0 and β_1 are unknown parameter

[1] Maarten Jansen is Professor at the Departments of Mathematics and Computer Science of the Université libre de Bruxelles (Belgium). He is Elected member of the International Statistical Institute and author of three books: *Wavelets from a statistical perspective* (CRC Press, 2022), *Second generation wavelets and applications* (with P. Oonincx, Springer Verlag, 2005) and *Noise reduction by wavelet thresholding* (Springer Verlag, 2001), and over 30 journal papers.

[2] Gerda Claeskens is Professor at ORStat, Faculty of Economics and Business and at the Leuven Statistics Research Center of the KU Leuven (Belgium). She is Elected member of the International Statistical Institute, fellow of the American Statistical Association (2019) and the Institute of Mathematical Statistics (2014), and recipient of the Noether Young Scholar Award (2004). She is the author of about 90 journal papers and of the book *Model selection and model averaging* (with N.L. Hjort, Cambridge University Press, 2008). Currently she is Associate editor of the *Journal of the American Statistical Association*, *TEST*, and *International Statistical Review*.

vectors and ε is often assumed to be a normal random variable with zero mean and an unknown variance σ^2. In nonparametric regression we do not specify the functional form for the conditional mean of Y and write the model as $Y = f(X) + \varepsilon$, where X may be random, or take fixed values. The terminology smoothing arises from the commonly made assumption for most methods (however, see the wavelet section below) that the unspecified function f is smooth.

Nonparametric estimation starts with choosing a basis which defines a space of functions. The function f is then approximated within this space by $\tilde{f}(x) = \sum_{j=1}^{J} \beta_j \psi_j(x)$. The basis functions may also depend on further parameters, specifying for example the location. These may be estimated or specified in advance. Fourier series are one example. Nonparametric estimation of f then proceeds with estimating the unknown parameters. Spaces of functions are often infinite dimensional, hence the number of basis functions to be used, J in the above sum, is a tuning parameter. The more basis functions taken, the better the approximation will be, in general. However, estimating more parameters comes at a cost of increased variance and increased computational effort.

The smoothing methods are in a similar way used for the estimation of density functions. While histograms give rough approximations, the nonparametric density estimators are smooth curves. For nonparametric density estimation, splines, wavelets or kernels may be used. For the latter method, see for example Wand and Jones (1995).

2.1 Spline estimation

The choice of the basis characterizes the estimated function. Often taken choices are *spline* functions. A jth degree polynomial spline is a curve that consists of piecewise jth degree polynomial parts that are continuously joined together at *knots*. The smoothness of the resulting function depends on whether also the higher derivatives of the spline are continuous. When each observation x_i, $i = 1, \ldots, n$ is taken as a knot, this results in a *smoothing spline*. When a set of knots $\kappa_1, \ldots, \kappa_K$ is chosen, with $K < n$, the sample size, the function is a *regression spline*. In cases that $K < n$, estimation of the unknown spline coefficients β_j can be done via least squares in case of (approximate) normal errors ε. For smoothing splines, one introduces a penalty term that is related to the derivatives of f. Also for regression splines, penalties on the coefficients may be stated, to reduce the influence of the choice of the knots. This results in *penalized regression splines*. An expanded description of spline regression methods can be found in the entry by Opsomer and Breidt (same volume). Some main references are Eubank (1988), Wahba (1990), Green and Silverman (1994) and Ruppert, Wand and Carroll (2003).

2.2 Wavelet estimation

Thanks to a fast decomposition algorithm (Mallat, 1989), wavelet bases have gained considerable success as a representation for data to be smoothed. Wavelet basis

functions are short waveforms located at a specific point in time or space and with a specific scale. This locality in time and frequency provides a tool for a multiscale and sparse representation of data. Especially piecewise smooth data, with isolated singularities, or data with otherwise intermittent behavior are typical objects for which wavelets are well suited. Indeed, the singularities can be captured by a relatively limited number of local wavelet basis functions, with appropriate scales, while the smooth intervals in between the singularities produce many but small contributions in a wavelet decomposition.

While other methods may have difficulties in catching singularities, in a wavelet decomposition they pose no bottleneck, provided that at the position of a singularity the wavelet representation is locally more refined than in between the singularities. The location of the singularities is done automatically, even in the presence of noise, by the fact that the coefficients corresponding to the basis functions at those positions are large, as they carry the contributions that constitute the singularity. Singularities are thus well captured by selecting the largest coefficients, rather than a predetermined subset. Putting the smallest coefficients to zero therefore removes most of the noise without affecting the noise-free data too much. The usage of this thresholding or any sophisticated variant, which is always a nonlinear processing, is the main argument for using a wavelet decomposition. The nonlinear processing (keeping in mind that the wavelet forward transform and reconstruction themselves are linear) relates directly to the intermittent nature of the data, i.e., the presence of isolated singularities in otherwise smooth behavior. The use of thresholds relies on the sparsity property of a wavelet respresentation. The multiscale property, on the other hand, is mostly used for additional across-scale processing, for instance to remove false positives after thresholding (for smoother intervals between singularities) or to correct for false negatives by looking across scales (for sharper reconstruction of singularities). Also scale dependent processing is necessary in the case of correlated noise on the observations (Donoho and Johnstone, 1995).

The selection of appropriate thresholds has been a major domain of research. Limiting or even reducing to zero the number of false positives is the objective of an important class of thresholds, including the universal threshold (Donoho and Johnstone, 1994) or False Discovery Rate thresholds (Benjamini and Hochberg, 1995). Another class of thresholds focusses on the expected, integrated squared loss, i.e., risk, of the result. Stein's Unbiased Risk Estimator and modifications (such as cross-validation) provide practical methods for finding minimum risk thresholds. A third, and wide class of threshold assessment methods is based on Bayesian — mostly empirical Bayes — models, such as EBayesthresh (Johnstone and Silverman, 2004, 2005). The prior model for noise-free coefficients reflects the idea of sparsity, mostly through a zero-inflated or otherwise mixture model with heavy tails (where heavy here includes everything heavier than the normal distribution).

2.3 Kernel and local polynomial estimation

Kernel estimation of a regression function starts from the idea that the function is locally well approximated by a low order polynomial curve. The Nadaraya-Watson estimator locally approximates the curve f at value x by a constant regression function. Observations X_i close to x get a large weight, and observations further away receive less or zero weight. The kernel function K determines the weighting and is assumed to be a density function. The estimator takes the following form, $\widehat{f}_h(x) = \sum_{i=1}^n K_h(x - X_i)Y_i / \sum_{i=1}^n K_h(x - X_i)$, where h is called the bandwidth. This is a tuning parameter, small values of h imply that only close neighbours get a large weight, this might result in a rather wiggly fit. Large values of h will result in much smoother fitted curves. Several studies have focussed on appropriate bandwidth choices, for example via cross-validation or plug-in methods based on asymptotic properties of the estimator. Variants on this estimator are the Priestley-Chao and Gasser-Müller estimator. Local polynomial estimators are similar in spirit. Instead of taking a local constant approximation of the function f around x, a local polynomial approximation is obtained. More information on kernel regression methods can be found in the entry by Opsomer and Breidt (same volume). For more details, see Fan and Gijbels (1996).

3 Other applications of nonparametric estimation

Nonparametric estimation is used beyond the classical regression models and density estimation as well. Examples of its use are found in functional data analysis and functional regression models where the response or some of the covariates, or both, are functions (Hsing and Eubank, 2015). A flexible modeling of generalized additive models makes use of nonparametric estimators, and in particular spline estimators, to estimate smooth functions for the location, scale and shape parameters of the wide class of distributions that fits in this framework (Rigby, et al., 2019).

References

[1] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, **57**:289–300.

[2] Eubank, R. (1988). *Spline Smoothing and Nonparametric Regression.* Marcel Dekker, New York.

[3] Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, **81**(3), 425–455.

[4] Donoho, D. and Johnstone, I. (1995). Adapting to unknown smoothness via wavelet-shrinkage. *J. Amer. Statist. Assoc.*, **90**(432):1200–1224.

[5] Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman and Hall.

[6] Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: A roughness penalty approach.* Chapman & Hall.

[7] Hsing, T. and Eubank, R. (2015). *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*, Wiley.

[8] Hollander, M. and Wolfe, D.A. (1999). *Nonparametric Statistical Methods*, 2nd Edition, Wiley.

[9] Johnstone, I. and Silverman, B. (2004). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics*, **32**(4), 1594–1649.

[10] Johnstone, I. and Silverman, B. (2005). Empirical bayes selection of wavelet thresholds. *Annals of Statistics*, **33**(4), 1700–1752.

[11] Mallat, S. (1989). A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**(7), 674–693.

[12] Rigby, R.A., Stasinopoulos, M.D., Heller, G.Z., and De Bastiani, F. (2019). *Distributions for Modeling Location, Scale, and Shape: Using GAMLSS in R.* Chapman & Hall/CRC. The R Series.

[13] Ruppert, D., Wand, M.P. and Carroll, R.J. (2003). *Semiparametric regression.* Cambridge University Press.

[14] Wahba, G. (1990). *Spline models for observational data.* SIAM.

[15] Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing.* Chapman and Hall.

[16] Wolfowitz, J. (1942). Additive partition functions and a class of statistical hypotheses. *The Annals of Statistics*, **13**, 247–279.