

Some Themes in Nonparametric Statistics

Ronald H. Randles

August 6, 2001

SOME THEMES IN NONPARAMETRIC STATISTICS 1975-2000

RONALD H. RANDLES
UNIVERSITY OF FLORIDA

Slides and Handout are available at:
<http://web.stat.ufl.edu/~rrandles/NPSurvey>

Working Definition:

Nonparametric Methods

\equiv (Methods that are Strictly Parametric)^C

THEMES

- **EXPLORING DATA**
- **ROBUSTNESS**
- **BOOTSTRAP**
- **SMOOTHING**
- **QUANTILES**
- **RANK METHODS FOR LINEAR MODELS**
- **MULTIVARIATE METHODS**
- **Omitted due to lack of time:**
 - **RELIABILITY**
 - **METHODS FOR SPECIAL DESIGNS**
 - **BIostatISTICS AND SURVIVAL ANALYSIS METHODS**
 - **ALGORITHMIC MODELLING**
- **GENERAL TEXTS**

EXPLORING DATA

John Tukey's text "Exploratory Data Analysis" (1977) pointed the field of statistics and data analysis in new directions.

The texts "Understanding Robust and Exploratory Data Analysis" (1983) and "Fundamentals of Exploratory Analysis of Variance" (1991) edited by Hoaglin, Mosteller and Tukey, illustrate many of the most important ideas and tools of EDA with numerous enlightening examples.

ROBUSTNESS

TOOLS

Hampel (1968, 1974) defined the **influence function** of a statistical functional $T(F)$ as:

$$IF(T, x) = \lim_{\epsilon \rightarrow 0} \frac{T((1 - \epsilon)F + \epsilon\Delta_x) - T(F)}{\epsilon}.$$

Here Δ_x is the df of a distribution that puts probability 1 on the value x . The IF measures the change in the functional, of a small amount of contamination at the point x . A $T(F)$ with an IF that is bounded in x is more robust to extreme values. But the IF is also related to the efficiency since:

$$Asy.Var.(T(F_n)) = \int [IF(T, x)]^2 dF(x)$$

Hampel (1968, 1971) also defined the **breakdown point** of the estimator $T(F)$ as the smallest fraction of contamination that could take the functional to arbitrarily extreme values.

CLASSES OF ESTIMATORS

M-Estimates which satisfy

$$\sum_{i=1}^n \Psi(X_i - \hat{\theta}) = 0.$$

L-Estimates which are linear functions of the order statistics:

$$\hat{\theta} = \sum_{i=1}^n c_i X_{(i)}.$$

R-Estimates chosen to satisfy

$$\sum_{i=1}^n a(R_i(\hat{\theta})) = 0$$

where $R_i(\theta)$ is the rank of the residual of an observation X_i around a model involving θ .

Minimum Distance Estimates where $\hat{\theta}$ minimizes

$$D(F_n, G_\theta).$$

S-Estimates which minimize a scale statistic

$$S(X_1 - \hat{\theta}, \dots, X_n - \hat{\theta})$$

which satisfies

$$\frac{1}{n} \sum_{i=1}^n \rho\left(\frac{X_i - \hat{\theta}}{S}\right) = K.$$

- The texts by: Serfling (1980), Hampel, Ronchetti, Rousseeuw and Stahel (1986), Sen and Singer (1993), Hettmansperger and McKean (1998), and Lehmann (1999) provide properties and large sample approximations of the distributions of these broad classes of estimators.
- **Adaptive Procedures** that selected among members of one of these classes were described by Hogg (1974, 1976).

- Donoho and Huber (1983) have contributed toward popularizing the use of the breakdown point as a fundamental robustness property.
- Considerable research has gone into developing the influence functions and breakdown points for classes of estimators. For example, in the one sample location problem an M-estimator with a bounded, monotone Ψ will have a bounded influence function and a breakdown point of $1/2$. See Hampel (1971) and Huber (1981).
- For higher dimensional data, a robust M -estimator will have a bounded influence function, but the breakdown is typically $(1 + \textit{dimension})^{-1}$.

- Rousseeuw (1985) proposed multivariate location estimators with a breakdown that does not decrease with dimension. His minimum volume ellipsoid and minimum covariance determinant estimators have been widely cited.
- The text by Hampel, Ronchetti, Rousseeuw and Stahel (1986) contains considerable information on robust estimation and testing, including estimation of location and covariance matrices for multivariate data.
- A recent vignette by Portnoy and He (2000) also contains an interesting discussion of robust methods and numerous references.

BOOTSTRAP

Treats the sample like the population and draws samples from this pseudo population in order to assess:

- (a) variability of an estimator
- (b) bias of an estimator
- (c) predictive performance of a rule
- (d) significance of a test

It was introduced by Efron (1979) initially as a way to describe the utility and properties of the Jackknife.

The beauty of the bootstrap is in its:

- (i) wide applicability
- (ii) increased accuracy
- (iii) ability to take advantage of modern computing

Confidence Interval Construction

- The percentile interval was proposed by Efron in (1979).
- A more sophisticated bias corrected and accelerated interval was described in Efron (1987).
- The second order accuracy of a bootstrap interval was shown by Singh (1981) and Hall (1988).
- Work on coverage improvement includes Sheather (1987) and Hall and Martin (1988).

Examples of its use

Variability: Hutson and Ernst (2000)

Bias: Efron (1990)

Predictive Performance: Efron and Gong (1983)

Significance: Romano (1988)

Resources:

Texts by:

Hall (1992)

Efron and Tibshirani (1993)

Shao and Tu (1995)

Davison and Hinkley (1997)

Articles by:

Presnell and Hall(1999)

Efron (2000)

SMOOTHING - Regression

Provide a “smooth” functional relationship between x and y that is not parametrically structured (linear, quadratic, etc.).

Some of the early work on this topic were the papers by Nadaraya (1964) and Watson (1964).

Kernel Smoothers:

$$s(x_0) = \frac{\sum_{i=1}^n K\left(\frac{x_0 - x_i}{h}\right) y_i}{\sum_{i=1}^n K\left(\frac{x_0 - x_i}{h}\right)}$$

There are important issues associated with the use of kernel smoothers - choice of kernel and the smoothing parameters.

- Härdle, Hall and Marron (1992) is one paper which proposes bandwidth selectors with good asymptotic properties.
- Cleveland (1979) fits a locally linear kernel estimator of the function. This process is implemented in Loess in S. Also see Tibshirani and Hastie (1987) and Fan and Gijbels (1996).
- Recent textbooks by Wand and Jones (1995) and Bowman and Azzolini (1997) are good sources of information and references.
- Smoothing splines were popularized by Wahba and Wold (1975). The text by Eubank (1999) is a good reference for smoothing splines.
- Other approaches include:
 - Fourier series estimators
 - Wavlets
- The texts by Simonoff (1996) and Hart (1997) describe the variety of smoothing approaches available and provide good references.

SMOOTHING - Density Estimation

Example of a kernel smoothed density estimate.

Kernel density estimates were described in a technical report by Fix and Hodges (1951) with early important contributions made by Rosenblatt(1956) and Parzen(1962).

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right),$$

where $\int_{-\infty}^{\infty} K(t)dt = 1$.

Choice of the kernel and choice of the smoothing parameter are important aspects of the performance of the estimator.

- The texts by Silverman (1986), Härdle (1990) and Scott (1992) provide details on the theory and practice of these estimators, plus numerous references.
- Schucany (1989) for a discussion about use of local bandwidths, and
- Hall (1990) for variable kernel methodology.

QUANTILES

- Parzen (1979) introduced a quantile function approach to data analysis, unifying a number of objectives in statistical inference. He introduced the density quantile function and kernel quantile estimators.
- The monograph by Csörgö (1983) provides an extensive treatment of the limiting distribution properties of quantile processes and simultaneous confidence bounds for quantiles.
- Babu and Rao (1988) find the joint asymptotic distributions for the marginal quantiles when sampling from a multivariate population.
- Padgett (1986) created a kernel quantile estimator for censored data.
- Sheather and Marron (1990) studied many different quantile estimation schemes, showing how many could be viewed as kernel estimators and that some are asymptotically equivalent. They also recommended choices for the smoothing parameters.
- Mudholkar and Hutson (1997) examine improved ways to estimate the sample quartiles.

LINEAR MODELS BASED ON RANK STATISTICS

- Jureckova (1971) proposed linear models estimators based on a rank statistic which correlated the (scored) ranks of the residuals with the independent variable.

Let $(Y_1, \mathbf{x}_1), \dots, (Y_n, \mathbf{x}_n)$ denote the observed pairs of dependent variable, Y_i , and vector valued independent variable, $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$.

Statistics:

$$L_j = \sum_{i=1}^n (x_{ij} - \bar{x}_{.j}) a(R_i \boldsymbol{\beta})$$

where $R_i \boldsymbol{\beta}$ is the rank of $Y_i - \boldsymbol{\beta}' \mathbf{x}_i$ among

$$(Y_1 - \boldsymbol{\beta}' \mathbf{x}_1), \dots, (Y_n - \boldsymbol{\beta}' \mathbf{x}_n).$$

and $a(i) = \phi\left(\frac{i}{(n+1)}\right)$.

- Puri and Sen (1985) give a very complete account of tests and estimators from such statistics, including:
 1. computation and asymptotics of estimators
 2. permutation principles and asymptotics of tests
 3. aligned rank tests for subhypotheses.

- Jaeckel (1972) proposed estimating β to minimize the dispersion criterion:

$$D_J(\beta) = \sum_{i=1}^n a(R_{i\beta}) * (Y_i - \beta' \mathbf{x}_i)$$

with, for example:

$$a(i) = \phi\left(\frac{i}{n+1}\right) \text{ and } \phi(u) = \sqrt{12}\left(u - \frac{1}{2}\right).$$

- Hettmansperger and McKean (1983) followed with tests of linear hypotheses based upon the reduction in dispersion (analogous to ANOVA).
- These procedures, which are robust in the response space, are implemented in MINITAB.
- See chapter 3 in Hettmansperger and McKean (1998) for additional details.

High Breakdown R-Linear Models

- Sievers (1983) and Naranjo and Hettmansperger (1994) develop a pseudo-norms which yield estimates $\hat{\beta}$ with bounded influence functions and positive or high breakdown properties in both the factor space and response space.
- See Simpson, Rupert and Carroll (1992), Chang, McKean, Narajo and Sheather (1999) and chapter 5 of Hettmansperger and McKean (1998) for different approaches to this problem.
- The vignette by Hettmansperger, McKean and Sheather (2000) provides many references.

Diagnostics

- McKean, Sheather and Hettmansperger (1990, 1993) explore the use of plots of the residuals from R-estimates (Jaeckel fits) to investigate lack of fit issues.
- They conclude, for example, that:
 1. Plots of residuals versus predicted values should be a random scatter when the model is correct.
 2. Plot of the residuals versus (the residual of a new independent variable(s)) should also be a random scatter if the new variable is not needed.
 3. Standardized R-residuals are useful in detecting outliers.
- Residuals from high-breakdown fits do not in general have the same diagnostic properties.

MULTIVARIATE METHODS

Componentwise Methods

- Puri and Sen (1971) classic textbook on rank based methods and their theory – focused primarily on univariate test statistics and estimators that are computed on the comparable components of vectors, which are then combined into an overall statistic
- Wegman (1986) – parallel coordinates graphs – to visualize the relationships among sets of multivariate data points
- Scale Invariant (Equivariant) Methods

- Chakraborty, Chaudhuri and Oja (1998) – describe a transformation-retransformation process of determining a multivariate extension of the sample median.
- Peters and Randles(1990) and Hössjer and Croux (1995) proposed multivariate extensions of the signed rank test.
- Randles(2000) – proposed a multivariate sign test with a distribution-free property.

Slides and Handout are available at:

<http://web.stat.ufl.edu/~rrandles/NPSurvey>