STAT 824 sp 2023 Lec 00 slides

What is nonparametric inference?

Karl B. Gregory

University of South Carolina

These slides are an instructional aid; their sole purpose is to display, during the lecture, definitions, plots, results, etc. which take too much time to write by hand on the blackboard. They are not intended to explain or expound on any material.

What is nonparametric inference?

Methods based on ranks form a substantial body of statistical techniques that provide alternatives to the classical parametric methods...the feature of nonparametric methods mainly responsible for their great popularity (and to which they owe their name) is the weak set of assumptions required for their validity.

– E.L. Lehmann in his 1975 book *Nonparametrics* [1]

The basic idea of nonparametric inference is to use data to infer an unknown quantity while making as few assumptions as possible. Usually, this means using statistical models that are infinite-dimensional.

– Larry Wasserman in his 2006 book *All of Nonparametric Statistics* [3]

The problem of nonparametric estimation consists in estimation, from the observations, of an unknown function belonging to a sufficiently large class of functions.

– Alexander Tsybakov in his 2008 book *Intro. to Nonparametric Estimation* [2].

Lehmann (1975) says "rank-based statistics constitute methodologically the most important part" of nonparametric statistics. Some of these are:

1. Wilcoxon rank-sum test for a difference in location.
2. Siegel-Tukey test for a difference in dispersion.
3. Sign test for paired comparisons.
4. Wilcoxon signed-rank test for paired comparisons.
5. The Kruskall-Wallace test for comparing several treatments.

Wasserman (2006) classes the above as "traditional" nonparametric methods.

Two central problems in "modern" nonparametrics

1. Density estimation: Let $X_1, \ldots, X_n$ iid with density f. Estimate f.
2. Regression function estimation: Let $Y_i = m(x_i) + \varepsilon_i$, $i = 1, \ldots, n$, $\varepsilon_1, \ldots, \varepsilon_n$ independent with mean zero and $x_1, \ldots, x_n$ real numbers. Estimate m.

Idea is to estimate f amd m while making very weak assumptions about them.

Spcf. assume they belong to classes of functions that are smooth in some sense.

Parametric estimation assumes f and m belong to parametric families, i.e.

1. $f \in \{g(x;\theta) : \theta \in \Theta\}$ where each $g(\cdot;\theta)$ is a known density and $\Theta \subset \mathbb{R}^d$.
2. $m \in \{g(x;\theta) : \theta \in \Theta\}$ where each $g(\cdot;\theta)$ is known and $\Theta \subset \mathbb{R}^d$.

The dimension d is finite and fixed (does not change with n).

Does the bootstrap belong to nonparametrics?

The bootstrap belongs to nonparametrics for traditional and modern reasons:

1. Traditional: Used when you do not want to make distributional assumptions about the data (*making as few assumptions as possible*).
2. Modern: Used to estimate unknown sampling distributions – often the quantiles of pivotal quantities (*estimation of an unknown function*).

Example:

- If $X_1, \ldots, X_n \overset{\text{ind}}{\sim} \text{Normal}(\mu, \sigma^2)$, then $\sqrt{n}(\bar{X}_n - \mu)/S_n \sim t_{n-1}$. Therefore

$$\bar{X}_n \pm t_{n-1,\alpha/2} S_n/\sqrt{n} \text{ is a } (1-\alpha) \times 100\% \text{ CI for } \mu.$$

- If $X_1, \ldots, X_n$ iid w/mean μ, var σ^2, and cdf F, then $\sqrt{n}(\bar{X}_n - \mu)/S_n \sim \mathcal{G}_n$.

$$(\bar{X}_n - \mathcal{G}_{n,\alpha/2} S_n/\sqrt{n}, \bar{X}_n - \mathcal{G}_{n,1-\alpha/2} S_n/\sqrt{n}) \text{ is a } (1-\alpha) \times 100\% \text{ CI for } \mu.$$

We use the bootstrap to estimate the quantiles $\mathcal{G}_{n,\alpha/2}$ and $\mathcal{G}_{n,1-\alpha/2}$ of $\mathcal{G}_n$.

Overview of STAT 824 topics

1. Density estimation: Estimating the cdf, kernel density estimation, Lipschitz and Hölder classes of densities, bounds on the MSE of kernel density estimators, bounds on the mean integrated squared error (MISE) of kernel density estimators, multivariate kernel density estimation, the "curse of dimensionality".
2. Nonparametric regression: Nadaraya-Watson estimator, MSE bounds under Lipschitz and Hölder smoothness, local polynomial estimator, least-squares splines, penalized splines and trend filtering, additive model, sparse additive model.
3. Bootstrap: Bootstrap for the mean, Edgeworth expansion and second-order correctness of the bootstrap, bootstrap for statistical functionals, bootstrap for linear regression, bootstrap in nonparametric regression.
4. Rank-based methods: Wilcoxon rank-sum test, some asymptotics for rank-based tests.


E L Lehmann.
Nonparametrics.
Holden-Day, Inc., 1975.


Alexandre B Tsybakov.
Introduction to nonparametric estimation.
Springer Science & Business Media, 2008.


Larry Wasserman.
All of nonparametric statistics.
Springer Science & Business Media, 2006.