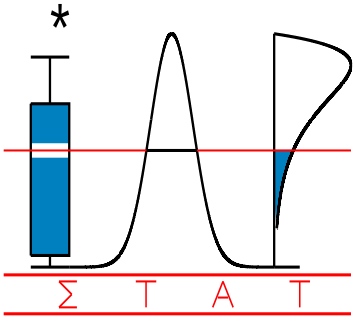
T E C H N I C A L

R E P O R T

0613

INVARIANCE, SEMIPARAMETRIC EFFICIENCY, AND RANKS

HALLIN M.



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

http://www.stat.ucl.ac.be/IAP

Happy Birthday* to you, Mr Wilcoxon!

Invariance, Semiparametric Efficiency, and Ranks

Marc Hallin**

Département de Mathématique, Institut de Statistique and ECARES
Université libre de Bruxelles Campus de la Plaine CP210
B-1050 Bruxelles, Belgium
mhallin@ulb.ac.be

1 Introduction

Everybody who went through a statistics course, even at introductory level, has been exposed at least to some elementary aspects of rank-based methods, and has heard about Wilcoxon's *signed rank* and *rank sum* tests.

Although early ideas of distribution-free tests can be traced back as far as John Arbuthnoth (1667-1735), Frank Wilcoxon's pathbreaking 1945 four page paper (Wilcoxon 1945), where these two tests are described for the first time, certainly can be considered as the starting point of the modern theory of rank-based inference.

The Wilcoxon tests, and most of the subsequent theory of rank-based inference, were developed as a reaction against the pervasive presence of Gaussian assumptions in classical statistical theory. Rank tests are simple and easy to compute. Above all, they are distribution-free, hence exact and applicable under unspecified (typically, non Gaussian) densities. Moreover, they are flexible enough to adapt to a wide range of inference problems: the fifties and the sixties have witnessed an explosive but somewhat aphazard development of rank-based solutions to a variety of problems. This development was structured and systematized in the seventies, mainly on the basis of Jaroslav Hájek's fundamental contribution, leading to what can be considered as the "classical theory" of rank-based inference. This classical theory essentially addresses all testing (and estimation) problems arising in the context of general linear models with independent observations, thus covering location, scale, and regression problems, as well as analysis of variance and covariance, and

* The initial version of this conference was delivered in Leuven as the inaugural lecture of a Francqui Chair on October 20, 2005—thus plainly justifying the title.

** Research supported by an I.A.P. contract of the Belgian Federal Government and an Action de Recherche Concertée of the Communauté française de Belgique. Special thanks are due to Davy Paindaveine for his careful reading of the manuscript and helpful comments.

most linear experimental planning models—see the monograph by Puri and Sen (1985) for a systematic and fairly exhaustive account.

The progress since then may have been less spectacular, and the opinion is not uncommon that rank-based inference is a more or less complete—hence limited and somewhat old-fashioned—theory, the development of which has stopped in the early eighties. The objective of this nontechnical presentation is to dispel this wrong perception by showing that, quite on the contrary, ranks and their generalizations quite naturally find their ultimate expression in the modern theories of asymptotic statistical experiments and semiparametric inference. More precisely, rank-based methods, under very general assumptions, allow for semiparametrically efficient, yet distribution-free inference (testing and estimation), in a very large variety of models involving unspecified densities—much beyond the classical linear models with independent observations.

Frank Wilcoxon himself would be most surprised to see how, in slightly more than sixty years, his two tests, which he modestly considered as quick and easy tricks, to be used when everything else fails, not only have survived the many revolutions of contemporary statistics, but have turned into a timely and still growing area of modern inference, reconciling the irreconcilable objectives of efficiency and robustness.

After sixty years of unremitting service, not the slightest prospect of early retirement, thus: happy birthday to you, Mr Wilcoxon!

2 Ranks: from distribution-freeness to group invariance

2.1 Ranks and rank tests

Let us first introduce some basic concepts and notation. Denoting by $\mathbf{X}^{(n)} := (X_1,\ X_2,\ \ldots,\ X_n)$ an n-tuple of observations, the *order statistic* is obtained by ordering the X_i's from smallest to largest: $\mathbf{X}_{(\ )} := (X_{\min} := X_{(1)}, X_{(2)}, \ldots X_{(n)} =: X_{\max})$, with $X_{(1)} \leq X_{(2)}, \leq \ldots \leq X_{(n)}$. The vector of ranks then is defined as $\mathbf{R}^{(n)} := (R_1,\ R_2,\ \ldots,\ R_n)$, with R_i such that $X_{(R_i)} = X_i$ or, equivalently, $R_i := \#\{j \mid X_j \leq X_i\}$.

This vector $\mathbf{R}^{(n)}$, provided that no ties occur (which happens with probability one as soon as $\mathbf{X}^{(n)}$ has a density), clearly is a (random) permutation of $(1,\ 2,\ \ldots,\ n)$. Assuming furthermore that the X_i's are i.i.d., with some unspecified density f over $\mathbb{R}$, the distribution of $\mathbf{R}^{(n)}$ is uniform over the $n!$ permutations of $(1, ..., n)$. An important advantage of $\mathbf{R}^{(n)}$-measurable statistics over the more general $\mathbf{X}^{(n)}$-measurable ones is thus their *distribution-freeness*: since their distributions do not depend on f, they allow for exact inference, robust to misspecification of f (hence to violations of Gaussian assumptions).

The price to be paid for this advantage is the corresponding loss of information. The observation $\mathbf{X}^{(n)}$ and the couple $(\mathbf{X}_{(\ )}, \mathbf{R}^{(n)})$ contain the same information: restricting to rank-based inference thus means throwing away

the information contained in the order statistic $\mathbf{X}_{(\ )}$. Natural questions are: how crucial (in terms of efficiency) is that loss of information? what is the real cost of this conflict between robustness and efficiency? The surprising answer (an answer Wilcoxon definitely would never have dreamed of) is: in case the density f is unknown, the loss of information is nil (asymptotically), and robustness can be obtained at no cost!

Fig. 1. Frank Wilcoxon (1892-1965)

The Wilcoxon rank sum test addresses the same two-sample location problem as the classical two-sample Student test, the only difference being that the latter requires Gaussian densities. Under the null hypothesis $\mathcal{H}_0$, $X_1, \ldots, X_m, X_{m+1}, \ldots, X_n$ are i.i.d., with unspecified (nonvanishing) density f, whereas, under the alternative $\mathcal{H}_1$, i.i.d.-ness is the property of $X_1, \ldots, X_m, X_{m+1} - \theta, \ldots, X_n - \theta$, for some $\theta > 0$.

The Wilcoxon test statistic can be written as $S_W^{(n)} := \sum_{i=m+1}^{n} R_i$; unlike the Student test, the wilcoxon test does not require any assumption on the density f, and thus resists light as well as heavy tails. Wilcoxon himself in 1945 hardly realized the consequences and importance of his discovery: he mainly considered his test as a robust, "quick and easy" solution for the location shift problem—nothing powerful, though—to be used when everything else fails.

2.2 Hodges-Lehmann and Chernoff-Savage

Eleven years after Wilcoxon's seminal paper, a totally unexpected result was published by Hodges and Lehmann (1956). This result, which came as a shock to the statistical community, is the following:

$$\inf_f \mathrm{ARE}_f\,(\text{Wilcoxon / Student}) = .864\ .$$

Recall that the ARE (asymptotic Relative Efficiency) of a sequence ($\phi_1^{(n)}$, say) of statistical procedures with respect to another one ($\phi_2^{(n)}$, say) is the limit $\mathrm{ARE}_f(\phi_1^{(n)}\ /\ \phi_2^{(n)})$, when it exists, as $n \to \infty$, of the ratio $n_2(n)/n$ of the number $n_2(n)$ of observations it takes for $\phi_2^{(n_2(n))}$ to achieve the same performance as $\phi_1^{(n)}$.

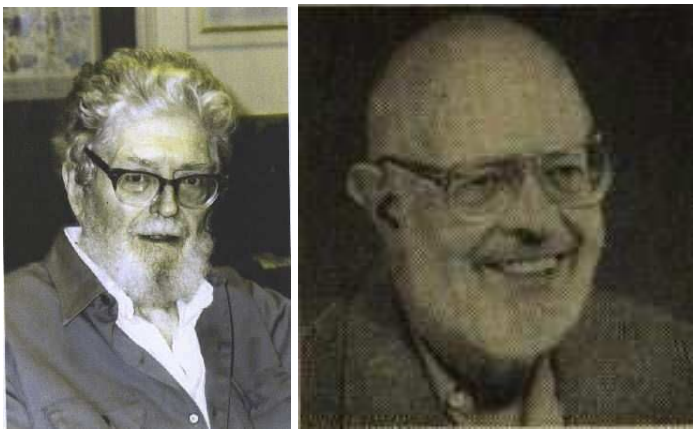


Fig. 2. Joseph L. Hodges (1922-2000) and Erich L. Lehmann (1917- —)

In the worst case, the Wilcoxon test thus only requires 13.6% more observations than the Student one in order to reach comparable power! On the other hand,

$$\sup_f \mathrm{ARE}_f\,(\text{Wilcoxon / Student}) = \infty,$$

and the benefits of unrestricted validity are invaluable ...

Since the Normal distribution is playing such a central role in classical statistics, the idea of considering, for the same location problem, a statistic of the form $S_{vdW}^{(n)} := \sum_{i=m+1}^{n} \Phi^{-1}\left(\frac{R_i}{n+1}\right)$ (or an equivalent *exact score* form), where Φ^{-1} denotes the standard normal quantile function, was proposed by several authors, among which Fisher, Terry, Yates, Fraser, van der Waerden, ... For simplicity, we all call them *van der Waerden statistics*.

Van der Waerden statistics are still distribution-free (since a function of ranks). In case however the actual underlying density is normal, $S_{vdW}^{(n)}$ is asymptotically equivalent to the Student statistic. Hence, at the normal, $S_{vdW}^{(n)}$ yields the same asymptotic performance as Student, which in that case is optimal.

Chernoff and Savage in 1958 (Chernoff and Savage 1958) however established the following much stronger result, which perhaps is even more surprising than Hodges and Lehmann's:

Fig. 3. Bartel L. van der Waerden (1903-1996)

$$\inf_f \mathrm{ARE}_f \text{ (van der Waerden / Student)} = 1.00 \ ,$$

an infimum which is attained at Gaussian f only!

It follows that van der Waerden tests are always strictly better (asymptotically) than the Student one, except at the normal, where they are equally good. One thus is always better off using van der Waerden which moreover, contrary to Student, is uniformly valid. This actually should put Student and much of everyday Gaussian practice out of business!

Fig. 4. Hermann Chernoff (1923- —) and I. Richard Savage (1926-2004)

.

These surprising facts cannot be a mere coincidence, and raise some obvious question: what is it that makes ranks that efficient? are ranks the only statistical objects enjoying such attractive distribution-freeness/efficiency prop-

erties? Answers however are not straightforward. As we shall see, they are intimately related with the maximal invariance property of ranks with respect to certain generating groups, and the connection of such invariance with tangent space projections and semiparametric efficiency. Such answers certainly were not at hand in 1958, and only emerged quite recently (Hallin and Werker 2003).

2.3 Group invariance

Assume that $\mathbf{X}^{(n)} = (X_1,\ X_2,\ \ldots,\ X_n)$ are i.i.d., with unspecified density f in the class $\mathcal{F}$ of all nonvanishing densities over $\mathbb{R}$ ($\mathbf{X}^{(n)}$ is thus *independent white noise*). Denote by $\mathrm{P}_f^{(n)}$ the joint distribution of $\mathbf{X}^{(n)}$ and let $\mathcal{P}^{(n)} := \{\mathrm{P}_f^{(n)} \mid f \in \mathcal{F}\}$.

Next consider the group of transformations (acting on $\mathbb{R}^n$)

$$\mathcal{G}, \circ := \left\{ g_h \mid h \text{ monotone } \uparrow,\ \text{continuous},\ h(\pm\infty) = \pm\infty \right\}, \circ$$

mapping $(x_1, \ldots, x_n) \in \mathbb{R}^n$ onto $g_h(x_1, \ldots, x_n) := (h(x_1), \ldots, h(x_n)) \in \mathbb{R}^n$. Then, $\mathcal{G}, \circ$ is a generating group for $\mathcal{P}^{(n)}$, in the sense that for all $\mathrm{P}_{f_1}^{(n)}$, $\mathrm{P}_{f_2}^{(n)}$ in $\mathcal{P}^{(n)}$, there exists $g_h \in \mathcal{G}$ such that $(X_1, \ldots, X_n) \sim \mathrm{P}_{f_1}^{(n)}$ iff $g_h(X_1, \ldots, X_n) \sim \mathrm{P}_{f_2}^{(n)}$. The vector of ranks $\mathbf{R}^{(n)}$ is maximal invariant for $\mathcal{G}, \circ$, that is, $T(x_1, \ldots, x_n) = T(g_h(x_1, \ldots, x_n))$ for all $g_h \in \mathcal{G}$ iff T is $\mathbf{R}^{(n)}$-measurable.

This invariance property of ranks suggests the definition of other "ranks", associated with other generating groups. Here are a few examples:

(i) $\mathbf{X}^{(n)} := (X_1,\ X_2,\ \ldots,\ X_n)$ i.i.d., with unspecified density f in the class $\mathcal{F}_+$ of all nonvanishing symmetric (w. r. t. 0) densities over $\mathbb{R}$ (*independent symmetric white noise*). Let $\mathcal{P}^{(n)} = \{\mathrm{P}_f^{(n)} \mid f \in \mathcal{F}_+\}$: this family is generated by the subgroup $\mathcal{G}_+, \circ$ of $\mathcal{G}, \circ$, where $\mathcal{G}_+ := \{g_h \in \mathcal{G} \mid h(-x) = h(x)\}$. The signs and the ranks of absolute values are maximal invariant ("signed ranks");

(ii) $\mathbf{X}^{(n)} := (X_1,\ X_2,\ \ldots,\ X_n)$ i.i.d., with unspecified nonvanishing median-centered density f in the class $\mathcal{F}_0$ of all nonvanishing zero-median densities over $\mathbb{R}$ (*independent median-centered white noise*). Let $\mathcal{P}^{(n)} = \{\mathrm{P}_f^{(n)} \mid f \in \mathcal{F}_0\}$: this family is generated by the subgroup $\mathcal{G}_0, \circ$ of $\mathcal{G}, \circ$, where $\mathcal{G}_0 := \{g_h \in \mathcal{G} \mid h(0) = 0\}$. The signs and the ranks are maximal invariant (see Hallin, Vermandele, and Werker 2006);

(iii) $\mathbf{X}^{(n)} := (X_1,\ X_2,\ \ldots,\ X_n)$ independent, with unspecified nonvanishing median-centered densities $f_1, \ldots, f_n$ in the class $\mathcal{F}_0$ of all nonvanishing zero-median densities over $\mathbb{R}$ (*independent, heterogeneous median-centered white noise*). Let $\mathcal{P}^{(n)} = \{\mathrm{P}_f^{(n)} \mid f \in \mathcal{F}_0\}$; the signs are maximal invariant for the appropriate generating group (see Dufour et al. 1998);

(iv) $\mathbf{X}^{(n)} := (\mathbf{X}_1,\ \mathbf{X}_2,\ \ldots,\ \mathbf{X}_n)$ i.i.d., with elliptical density

$$\sigma^{-k} (\det \mathbf{V})^{-1/2} f_1(\sigma^{-1} \sqrt{(\mathbf{x} - \boldsymbol{\mu})' \mathbf{V}^{-1} (\mathbf{x} - \boldsymbol{\mu})}).$$

over $\mathbb{R}^k$ (*independent elliptical white noise* with location $\boldsymbol{\mu}$, shape $\mathbf{V}$, scale σ, and standardized radial density f_1). Write $\mathbf{X} \sim \mathrm{P}^{(n)}_{\boldsymbol{\theta};f_1}$, $\boldsymbol{\theta} = (\boldsymbol{\mu}, \sigma, \mathbf{V})$ and $\mathcal{P}^{(n)} = \{\mathrm{P}^{(n)}_{\boldsymbol{\theta};f_1} \mid f_1 \in \mathcal{F}^+\}$, where $\mathcal{F}^+$ is the class of all standardized nonvanishing densities over $\mathbb{R}^+$: the unit vectors $\mathbf{U}_i := \mathbf{V}^{-1/2}(\mathbf{X}_i - \boldsymbol{\mu})/[(\mathbf{X}_i - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{X}_i - \boldsymbol{\mu})]^{1/2}$ and the ranks R_i of the "distances" $[(\mathbf{X}_i - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{X}_i - \boldsymbol{\mu})]^{1/2}$ are maximal invariant (multivariate signs $\mathbf{U}_i$ and ranks R_i) for the generating group of continuous order-preserving radial transformations: see Hallin and Paindaveine (2002 and 2006) for details.

It is easy to show that maximal invariants (hence, invariants) are distribution-free. As we shall see, they also have a strong connection to (semi-parametric) efficiency. This however requires some further preparation.

3 Efficiency: from parametric to semiparametric

3.1 Parametric optimality

In the sequel, we consider semiparametric models, namely, models under which the distribution of some $\mathcal{X}^n$-valued observation $\mathbf{X}^{(n)} := (X_1,\ X_2,\ \ldots,\ X_n)$ belongs to a family of the form $\mathcal{P}^{(n)} = \{\mathrm{P}^{(n)}_{\boldsymbol{\theta};f} \mid \boldsymbol{\theta} \in \boldsymbol{\Theta},\ f \in \mathcal{F}\}$ where $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^m$ is some m-dimensional parameter of interest, and $f \in \mathcal{F}$ is a nonparametric (infinite-dimensional) nuisance. We moreover assume that $\mathcal{P}^{(n)}$ is such that all its fixed-f parametric subfamilies $\mathcal{P}^{(n)}_f := \{\mathrm{P}^{(n)}_{\boldsymbol{\theta};f} \mid \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ are LAN (see below), whereas the fixed-$\boldsymbol{\theta}$ nonparametric subfamilies $\mathcal{P}^{(n)}_{\boldsymbol{\theta}} := \{\mathrm{P}^{(n)}_{\boldsymbol{\theta};f} \mid\ f \in \mathcal{F}\}$ are generated by some group $\mathcal{G}^{(n)}_{\boldsymbol{\theta}}, \circ$ acting on the observation space $\mathcal{X}^n$, with maximal invariant $\mathbf{R}^{(n)}(\boldsymbol{\theta})$.

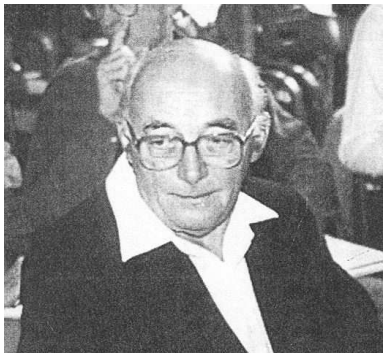


Fig. 5. Lucien Le Cam (1924-2000)

The concept of local asymptotic normality (LAN, w.r.t. $\boldsymbol{\theta}$, at given f) is due to Lucien Le Cam, and is now widely adopted as the standard structure

for traditional central-limit type asymptotics. The (sub)family $\mathcal{P}_f^{(n)}$ (more precisely, the sequence of families indexed by $n \in \mathbb{N}$) is said to be LAN if, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, there exists a random vector $\boldsymbol{\Delta}_{\boldsymbol{\theta};f}^{(n)}$ (the *central sequence*) and a (deterministic) positive definite matrix $\boldsymbol{\Gamma}_{\boldsymbol{\theta};f}$ (the *information matrix*) such that, under $\mathrm{P}_{\boldsymbol{\theta};f}^{(n)}$, as $n \to \infty$,

(i) $\Lambda_{\boldsymbol{\theta}+n^{-1/2}\boldsymbol{\tau}/\boldsymbol{\theta};f}^{(n)} := \log\left(\frac{\mathrm{dP}_{\boldsymbol{\theta}+n^{-1/2}\boldsymbol{\tau};f}^{(n)}}{\mathrm{dP}_{\boldsymbol{\theta};f}^{(n)}}\right) = \boldsymbol{\tau}'\boldsymbol{\Delta}_{\boldsymbol{\theta};f}^{(n)} - \frac{1}{2}\boldsymbol{\tau}'\boldsymbol{\Gamma}_{\boldsymbol{\theta};f}\boldsymbol{\tau} + o_{\mathrm{P}}(1)$, and

(ii) $\boldsymbol{\Delta}_{\boldsymbol{\theta};f}^{(n)} \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}_{\boldsymbol{\theta};f})$.

Skipping technical details, LAN implies that

(a) under $\mathrm{P}_{\boldsymbol{\theta}+n^{-1/2}\boldsymbol{\tau};f}^{(n)}$, $\boldsymbol{\tau} \in \mathbb{R}^m$, the central sequence $\boldsymbol{\Delta}_{\boldsymbol{\theta};f}^{(n)}$ is asymptotically $\mathcal{N}(\boldsymbol{\Gamma}_{\boldsymbol{\theta};f}\boldsymbol{\tau}, \boldsymbol{\Gamma}_{\boldsymbol{\theta};f})$ as $n \to \infty$;

(b) parametric efficiency (local, at $\boldsymbol{\theta}$, and asymptotic) in the initial (fixed-f) model has the same characteristics as parametric efficiency (exact) in the Gaussian shift model $\boldsymbol{\Delta} \sim \mathcal{N}(\boldsymbol{\Gamma}_{\boldsymbol{\theta};f}\boldsymbol{\tau}, \boldsymbol{\Gamma}_{\boldsymbol{\theta};f})$, $\boldsymbol{\tau} \in \mathbb{R}^m$, that is, for instance,
   - optimal α-level tests of $\mathcal{H}_0^{(n)} : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ achieve at $\mathrm{P}_{\boldsymbol{\theta}_0+n^{-1/2}\boldsymbol{\tau};f}^{(n)}$ asymptotic power $1 - F_{m;\boldsymbol{\tau}'\boldsymbol{\Gamma}_{\boldsymbol{\theta}_0;f}^{-1}\boldsymbol{\tau}}(\chi^2_{m;1-\alpha})$, where $F_{m;\lambda}$ stands for the noncentral chi-square distribution function with m degrees of freedom and noncentrality parameter λ, or
   - optimal estimates $\hat{\boldsymbol{\theta}}^{(n)}$ are such that $n^{1/2}(\hat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} \boldsymbol{\Gamma}_{\boldsymbol{\theta};f}^{-1}\boldsymbol{\Delta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}_{\boldsymbol{\theta};f}^{-1})$.

Moreover, optimality is achieved by treating the central sequence $\boldsymbol{\Delta}_{\boldsymbol{\theta};f}^{(n)}$ exactly as one would the observation $\boldsymbol{\Delta}$ in the limit Gaussian shift model, that is, for instance,

- by basing tests for $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ on the asymptotic χ^2_m null distribution of statistics of the form $Q_f := (\boldsymbol{\Delta}_{\boldsymbol{\theta}_0;f}^{(n)})'\boldsymbol{\Gamma}_{\boldsymbol{\theta}_0;f}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\theta}_0;f}^{(n)}$, or
- by constructing optimal estimators $\hat{\boldsymbol{\theta}}^{(n)}$ (of the *one-step form*) such that $n^{1/2}(\hat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}) = \boldsymbol{\Gamma}_{\boldsymbol{\theta};f}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\theta};f}^{(n)} + o_{\mathrm{P}}(1) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}_{\boldsymbol{\theta};f}^{-1})$.

Summing up, parametric efficiency (at given f and $\boldsymbol{\theta}$) is entirely characterized by the Gaussian shift model $\boldsymbol{\Delta} \sim \mathcal{N}(\boldsymbol{\Gamma}_{\boldsymbol{\theta};f}\boldsymbol{\tau}, \boldsymbol{\Gamma}_{\boldsymbol{\theta};f})$, $\boldsymbol{\tau} \in \mathbb{R}^m$, hence by the information matrix $\boldsymbol{\Gamma}_{\boldsymbol{\theta};f}$.

3.2 Parametric efficiency in the presence of nuisance

In order to understand what is meant with semiparametric efficiency, let us first consider the concept of parametric efficiency in the presence of a parametric nuisance. In the LAN family just described, assume that the parameter breaks into $\boldsymbol{\theta}' = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')$, and that inference is to be made about $\boldsymbol{\theta}_1 \in \mathbb{R}^{m_1}$, while $\boldsymbol{\theta}_2 \in \mathbb{R}^{m_2}$ is a nuisance. The central sequence $\boldsymbol{\Delta}_{\boldsymbol{\theta};f}^{(n)}$ similarly decomposes into $\begin{pmatrix} \boldsymbol{\Delta}_{\boldsymbol{\theta};f;1}^{(n)} \\ \boldsymbol{\Delta}_{\boldsymbol{\theta};f;2}^{(n)} \end{pmatrix}$, and the information matrix into $\boldsymbol{\Gamma}_{\boldsymbol{\theta};f} = \begin{pmatrix} \boldsymbol{\Gamma}_{\boldsymbol{\theta};f;11} & \boldsymbol{\Gamma}_{\boldsymbol{\theta};f;12} \\ \boldsymbol{\Gamma}_{\boldsymbol{\theta};f;12} & \boldsymbol{\Gamma}_{\boldsymbol{\theta};f;22} \end{pmatrix}$.

Inspired by exact optimality in the limit Gaussian shift, it is easy to understand that locally asymptotically optimal (efficient) inference on $\boldsymbol{\theta}_1$ should be based on the residual $\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f;1} - \boldsymbol{\Gamma}_{\boldsymbol{\theta};f;12}\boldsymbol{\Gamma}^{-1}_{\boldsymbol{\theta};f;22}\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f;2}$ of the regression of $\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f;1}$ on $\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f;2}$ in the covariance $\boldsymbol{\Gamma}_{\boldsymbol{\theta};f}$, that is, the $\boldsymbol{\Gamma}_{\boldsymbol{\theta};f}$-projection of the $\boldsymbol{\theta}_1$-central sequence parallel to the space of the $\boldsymbol{\theta}_2$-central sequence. Indeed, a local perturbation $n^{-1/2}\boldsymbol{\tau}_2$ of $\boldsymbol{\theta}_2$ induces (see (a) in Section 3.1) on the asymptotic distribution of $\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f}$ a shift $\begin{pmatrix}\boldsymbol{\Gamma}_{\boldsymbol{\theta};f;12}\\ \boldsymbol{\Gamma}_{\boldsymbol{\theta};f;22}\end{pmatrix}\boldsymbol{\tau}_2$. The resulting shift for the residual $\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f;1} - \boldsymbol{\Gamma}_{\boldsymbol{\theta};f;12}\boldsymbol{\Gamma}^{-1}_{\boldsymbol{\theta};f;22}\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f;2}$ is thus $\boldsymbol{\Gamma}_{\boldsymbol{\theta};f;12}\boldsymbol{\tau}_2 - \boldsymbol{\Gamma}_{\boldsymbol{\theta};f;12}\boldsymbol{\Gamma}^{-1}_{\boldsymbol{\theta};f;22}\boldsymbol{\Gamma}_{\boldsymbol{\theta};f;22}\boldsymbol{\tau}_2 = \mathbf{0}$: this residual therefore is insensitive to local perturbations of $\boldsymbol{\theta}_2$. On the other hand, the asymptotic covariance of the same the residual $\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f;1} - \boldsymbol{\Gamma}_{\boldsymbol{\theta};f;12}\boldsymbol{\Gamma}^{-1}_{\boldsymbol{\theta};f;22}\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f;2}$ is $\boldsymbol{\Gamma}_{\boldsymbol{\theta};f;11} - \boldsymbol{\Gamma}_{\boldsymbol{\theta};f;12}\boldsymbol{\Gamma}^{-1}_{\boldsymbol{\theta};f;22}\boldsymbol{\Gamma}'_{\boldsymbol{\theta};f;12}$, whereas a perturbation $n^{-1/2}\boldsymbol{\tau}_1$ of $\boldsymbol{\theta}_1$ induces a shift $\left(\boldsymbol{\Gamma}_{\boldsymbol{\theta};f;11} - \boldsymbol{\Gamma}_{\boldsymbol{\theta};f;12}\boldsymbol{\Gamma}^{-1}_{\boldsymbol{\theta};f;22}\boldsymbol{\Gamma}'_{\boldsymbol{\theta};f;12}\right)\boldsymbol{\tau}_1$. Asymptotically efficient (at given f and $\boldsymbol{\theta}$) inference on $\boldsymbol{\theta}_1$ when $\boldsymbol{\theta}_2$ is a nuisance is characterized by the Gaussian shift model

$$\boldsymbol{\Delta} \sim \mathcal{N}\left(\left(\boldsymbol{\Gamma}_{\boldsymbol{\theta};f;11} - \boldsymbol{\Gamma}_{\boldsymbol{\theta};f;12}\boldsymbol{\Gamma}^{-1}_{\boldsymbol{\theta};f;22}\boldsymbol{\Gamma}'_{\boldsymbol{\theta};f;12}\right)\boldsymbol{\tau}, \boldsymbol{\Gamma}_{\boldsymbol{\theta};f;11} - \boldsymbol{\Gamma}_{\boldsymbol{\theta};f;12}\boldsymbol{\Gamma}^{-1}_{\boldsymbol{\theta};f;22}\boldsymbol{\Gamma}'_{\boldsymbol{\theta};f;12}\right),$$

$\boldsymbol{\tau} \in \mathbb{R}^{m_1}$ hence by the information matrix $\boldsymbol{\Gamma}_{\boldsymbol{\theta};f;11} - \boldsymbol{\Gamma}_{\boldsymbol{\theta};f;12}\boldsymbol{\Gamma}^{-1}_{\boldsymbol{\theta};f;22}\boldsymbol{\Gamma}'_{\boldsymbol{\theta};f;12}$.

3.3 Semiparametric efficiency

In the previous two sections, the density f was supposed to be correctly specified. In a semiparametric context, of course, this density f is the nuisance, playing the role of $\boldsymbol{\theta}_2$! Except for the technical details related to the infinite-dimensional nature of f (the classical reference is the monograph by Bickel et al. 1993), this nuisance intuitively is treated in the same way as the parametric nuisance $\boldsymbol{\theta}_2$ in Section 3.2. Instead of being projected along the space of shifts induced by local variations of $\boldsymbol{\theta}_2$, however, $\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f}$ is projected along the space generated by the shifts induced by variations of densities in the *vicinity* of f: the so-called *tangent space*. This projected *semiparametrically efficient central sequence* $\boldsymbol{\Delta}^{(n)*}_{\boldsymbol{\theta};f}$, with (asymptotic) covariance $\boldsymbol{\Gamma}^{*}_{\boldsymbol{\theta};f} \leq \boldsymbol{\Gamma}_{\boldsymbol{\theta};f}$—the *semiparametrically efficient information matrix* in turn defines a Gaussian shift model $\boldsymbol{\Delta}^{*} \sim \mathcal{N}\left(\boldsymbol{\Gamma}^{*}_{\boldsymbol{\theta};f}\boldsymbol{\tau}, \boldsymbol{\Gamma}^{*}_{\boldsymbol{\theta};f}\right)$, $\boldsymbol{\tau} \in \mathbb{R}^{m}$ which characterizes the best performance that can be expected (at f and $\boldsymbol{\theta}$) when f is unspecified.

In some models, the semiparametric information matrix $\boldsymbol{\Gamma}^{*}_{\boldsymbol{\theta};f}$ coincides with the parametric one $\boldsymbol{\Gamma}_{\boldsymbol{\theta};f}$: the model is *adaptive at* f, meaning that parametric and semiparametric performances are asymptotically the same at f (possibly, at all f). In general, however, $\boldsymbol{\Gamma}^{*}_{\boldsymbol{\theta};f} < \boldsymbol{\Gamma}_{\boldsymbol{\theta};f}$: the cost of not knowing the *true density*, at f, is strictly positive.

Although the definitions of the semiparametrically efficient (at given f) central sequence and information matrix are intuitively satisfactory, their practical value at first sight is less obvious. While $\boldsymbol{\Gamma}^{*}_{\boldsymbol{\theta};f} \leq \boldsymbol{\Gamma}_{\boldsymbol{\theta};f}$ provides the

optimality bounds that in principle can be achieved at f, $\boldsymbol{\Delta}^{(n)*}_{\boldsymbol{\theta};f}$ heavily depend on f, and cannot be computed from the observations: $\boldsymbol{\Delta}^{(n)*}_{\boldsymbol{\theta};f}$ thus cannot be used for achieving the bound. This problem can be solved in two ways (recall that the central sequence at f—hence also the semiparametrically efficient one—only are defined up to $o_{\mathrm{P}^{(n)}_{\boldsymbol{\theta};f}}(1)$ terms).

(i) for all f in some class $\mathcal{F}$ of densities, an estimate $\hat{f}^{(n)}$ can be constructed in such a way that $\boldsymbol{\Delta}^{(n)*}_{\boldsymbol{\theta};\hat{f}^{(n)}} - \boldsymbol{\Delta}^{(n)*}_{\boldsymbol{\theta};f}$ under $P^{(n)}_{\boldsymbol{\theta};f}$ is $o_{\mathrm{P}}(1)$ as $n \to \infty$. Then, $\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta}} := \boldsymbol{\Delta}^{(n)*}_{\boldsymbol{\theta};\hat{f}^{(n)}}$, which is a measurable function of the observations, is asymptotically equivalent to the actual efficient central sequence for any $f \in \mathcal{F}$; together with $\boldsymbol{\Gamma}^{*}_{\boldsymbol{\theta}^{(n)}} := \boldsymbol{\Gamma}^{*}_{\boldsymbol{\theta};\hat{f}^{(n)}}$, it allows for uniformly (over $\mathcal{F}$) semiparametrically efficient inference. The convergence of the distribution of $\boldsymbol{\Delta}^{(n)*}_{\boldsymbol{\theta}}$ to a $\mathcal{N}\left(\mathbf{0}, \boldsymbol{\Gamma}^{*}_{\boldsymbol{\theta};f}\right)$ one, however, may be quite slow, and unpleasant technicalities such as *sample splitting* are often required.

(ii) if, for some selected f, a distribution-free statistic $\underset{\sim}{\boldsymbol{\Delta}}^{(n)}_{\boldsymbol{\theta};f}$ can be constructed such that $\underset{\sim}{\boldsymbol{\Delta}}^{(n)}_{\boldsymbol{\theta};f} - \boldsymbol{\Delta}^{(n)*}_{\boldsymbol{\theta};f}$ under $P^{(n)}_{\boldsymbol{\theta};f}$ is $o_{\mathrm{P}}(1)$ as $n \to \infty$, then this $\underset{\sim}{\boldsymbol{\Delta}}^{(n)}_{\boldsymbol{\theta};f}$ is a version of the semiparametrically efficient central sequence at f enjoying the remarkable property of being distribution-free, hence asymptotically $\mathcal{N}\left(\mathbf{0}, \boldsymbol{\Gamma}^{*}_{\boldsymbol{\theta};f}\right)$ irrespective of the actual underlying density, thus allowing for reaching semiparametric optimality at the selected f based on exact (even under density $g \neq f$) inference. As we shall see in the next section, this is precisely what rank-based inference can provide.

4 Ranks: from tangent space to Hájek projection

A fundamental statistical principle is the Invariance Principle, stipulating that "when a statistical problem is invariant under the action of some group of transformations, one should restrict to invariant statistical procedures", that is, to statistical procedures based on invariant statistics. It has been assumed in Section 3.1 that the fixed-$\boldsymbol{\theta}$ subfamilies $\mathcal{P}^{(n)}_{\boldsymbol{\theta}}$ of $\mathcal{P}^{(n)}$ are invariant w.r.t. the groups $\mathcal{G}_{\boldsymbol{\theta}}, \circ$, with maximal invariant $\mathbf{R}^{(n)}(\boldsymbol{\theta})$ (typically, the ranks of some $\boldsymbol{\theta}$-residuals). The set of invariant statistics thus coincides with the set of $\mathbf{R}^{(n)}(\boldsymbol{\theta})$-measurable statistics (typically, the rank statistics). Since optimal (at $\boldsymbol{\theta}$ and f) inference can be based on the central sequence $\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f}$, a natural idea consists in considering the invariant statistic which is closest to the central sequence by projecting $\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f}$ onto the σ-field generated by $\mathbf{R}^{(n)}(\boldsymbol{\theta})$, yielding

$$\underset{\sim}{\boldsymbol{\Delta}}^{(n)}_{\boldsymbol{\theta};f} := \mathrm{E}_f\left[\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f} \mid \mathbf{R}^{(n)}(\boldsymbol{\theta})\right]$$

Being $\mathbf{R}^{(n)}(\boldsymbol{\theta})$-measurable, $\underset{\sim}{\boldsymbol{\Delta}}^{(n)}_{\boldsymbol{\theta};f}$ is an invariant, hence distribution-free statistic (in the fixed-$\boldsymbol{\theta}$ submodel). The projection mapping $\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f}$ onto $\underset{\sim}{\boldsymbol{\Delta}}^{(n)}_{\boldsymbol{\theta};f}$ is, in a

sense, the opposite of a classical "Hájek projection"; in the sequel, as a tribute to Jaroslav Hájek, we also call it a *Hájek projection.*

The relation between the seemingly completely unrelated Hájek and tangent space projections was established by (Hallin and Werker 2003). Under very general conditions, indeed, they show that, under $\mathrm{P}^{(n)}_{\boldsymbol{\theta};f}$, $\underset{\sim}{\boldsymbol{\Delta}}{}^{(n)}_{\boldsymbol{\theta};f} = \boldsymbol{\Delta}^{(n)*}_{\boldsymbol{\theta};f} + o_{\mathrm{P}}(1)$ as $n \to \infty$: $\underset{\sim}{\boldsymbol{\Delta}}{}^{(n)}_{\boldsymbol{\theta};f}$ is thus an invariant (rank-based) distribution-free version of the semiparametrically efficient (at $\boldsymbol{\theta}$ and f) central sequence. As explained in Section 3.2, it thus allows for distribution-free semiparametrically efficient (at $\boldsymbol{\theta}$ and f) inference on $\boldsymbol{\theta}$.

Fig. 6. Jaroslav Hájek (1926-1974)

Remark that $\underset{\sim}{\boldsymbol{\Delta}}{}^{(n)}_{\boldsymbol{\theta};f}$ is obtained as the projection of the "regular" central sequence, not the semiparametrically efficient one: Hájek projections thus are doing the same job as tangent space projections, without requiring the (often nontrivial) computation of the latter, and with the (invaluable) additional advantages of distribution-freeness. The projection $\mathrm{E}_f[\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f} \mid \mathbf{R}^{(n)}(\boldsymbol{\theta})]$ is the "exact score version" of $\underset{\sim}{\boldsymbol{\Delta}}{}^{(n)}_{\boldsymbol{\theta};f}$; simpler "approximate score" versions also exist, but their form depends on the specific central sequence under study.

Uniformly semiparametrically efficient inference is also possible, by considering $\underset{\sim}{\boldsymbol{\Delta}}{}^{(n)}_{\boldsymbol{\theta};\hat{f}^{(n)}}$, where $\hat{f}^{(n)}$ is an appropriate density estimator, with the important advantage of avoiding the unpleasant technicalities, such as sample-splitting, associated with the "classical semiparametric procedures", based on $\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};\hat{f}^{(n)}}$. But then, $\underset{\sim}{\boldsymbol{\Delta}}{}^{(n)}_{\boldsymbol{\theta};\hat{f}^{(n)}}$ also splits the sample, into two mutually independent parts: the invariant and distribution-free part on one hand (the ranks), the "order statistic" (involved in $\hat{f}^{(n)}$ on the other, with the ranks containing the "f-free" information about the parameter $\boldsymbol{\theta}$, whereas the "order statistic" contains information on the nuisance f only.

5 Conclusion

Rank-based methods (more generally, the "maximal invariant" ones) are quite flexible, and apply in a very broad class of statistical models, much beyond the traditional context of linear models with independent observations. They are powerful—achieving semiparametric efficiency at selected density, which is the best that can be hoped for in presence of unspecified densities. In the same time, they are simpler and more robust (distribution-freeness) than "classical" semiparametric procedures. Often, they make Gaussian or pseudo-Gaussian methods non-admissible (the Chernoff-Savage phenomenon: see Hallin 1994 for time series models, Hallin and Paindaveine 2002 for elliptical location, and Paindaveine 2006 for elliptical shape).

Within their sixty years of existence, Wilcoxon's "quick and easy" tricks have grown into a full body of efficient and modern methods, reconciling the apparently antagonistic objectives of efficiency and robustness (distribution-freeness, meaning 100% resistance against misspecified densities).

Happy birthday to you, Mr Wilcoxon!

References

1. Bickel PJ, Klaassen CAJ, Ritov Y, Wellner JA (1993) Efficient and Adaptive Statistical Inference for Semiparametric Models. Johns Hopkins University Press, Baltimore.
2. Chernoff H, Savage IR (1958) Asymptotic normality and efficiency of certain nonparametric tests. Annals of Mathematical Statististics 29:972–994
3. Dufour JM, Hallin M, Mizera I (1998) Generalized run tests for heteroscedastic time series. Journal of Nonparametric Statistics 9:39–86
4. Hallin M (1994) On the Pitman nonadmissibility of correlogram-based time series methods. Journal of Time Series Analysis 16:607–612
5. Hallin M, Paindaveine D. (2002) Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks. Annals of Statistics 30:1103–1133
6. Hallin M, Paindaveine D. (2006) Semiparametrically efficient rank-based inference for shape: I Optimal rank-based tests for sphericity. Annals of Statistics 34
7. Hallin M, Vermandele C, Werker BJM (2006) Linear serial and nonserial sign-and-rank statistics: asymptotic representation and asymptotic normality. Annals of Statistics 34
8. Hallin M, Werker BJM (2003) Semiparametric efficiency, distribution-freeness, and invariance. Bernoulli 9:137–165
9. Hodges JL, Lehmann EL (1956) The efficiency of some nonparametric competitors of the t-test. Annals of Mathematical Statististics 27:324–335
10. Paindaveine D. (2006) A Chernoff-Savage result for shape. On the non-admissibility of pseudo-Gaussian methods. Journal of Multivariate Analysis: to appear
11. Puri ML, Sen PK (1985) Nonparametric Methods in General Linear Models, John Wiley New York
12. Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics Bulletin 1: 80–83