From Distribution-Freeness to Semiparametric Efficiency: Sixty(-three) Years of Rank-Based Inference

Marc Hallin

ECARES – Université Libre de Bruxelles
and European University Institute, Florence

Einaudi Institute for Economics and Finance, March 2008

Happy Birthday to You Mr Wilcoxon!

1945-2005 : 60th anniversary of

Wilcoxon (1945). Individual Comparisons by Ranking Methods,

Biometrics Bulletin **1**, 80-83

Wilcoxon 1945

- is only 3 pages, one line and 5 references long, with two pages developing numerical applications
- introduces Wilcoxon's rank sum and signed rank tests
- as rank-based alternatives to the traditional one- and two-sample Student tests for location

... and can be considered as the starting point of the modern history of rank-based inference

before Wilcoxon ...

Of course, there had been rank-based methods before 1945, some of which had a decisive influence upon Wilcoxon:

before Wilcoxon ...

● John Arbuthnot in 1710 has proposed the first *sign test*, which is also, probably the first hypothesis testing procedure ever

● Spearman as early as 1904 had introduced what is known as the Spearman rank correlation coefficient ρ

● Hotelling and Pabst (1936), revisiting Spearman's idea, had established the asymptotic normality of ρ (they even provide what we would call its ARE with respect to "parametric correlation" under the normal), whereas

● Friedman, in an ingenious 1937 work had shown how the comparison of several rankings (one in each block) allowed for testing the absence of treatment in two-way analysis of variance

● this line of research was pursued further by Kendall (1938), with his coefficient of correlation τ

● Wald and Wolfowitz (1943) extended the idea to a serial (time-series) context, also providing conditions for asymptotic normality, and

... *and Stigler's law of eponymy ...*

moreover, Stigler (1980)'s unyielding *Law of Eponymy* once again is strikingly confirmed :

● Spearman's ρ apparently is already present in Binet and Henry (1898)

● Kendall's *tau* similarly is a rediscovery of Lipps (1906), and it seems that

● Wilcoxon's two-sample test itself could have been found in Deuchler (1914) ...

... but these early contributions were largely ignored and had not the slightest impact on subsequent developments ...

...

● all results previous to Wilcoxon however are dealing, essentially, with comparing two or several rankings (the so-called m-rankings problem; cfr Spearman or Kendall)

which, albeit almost by accident also allows for testing against correlation, hence against autocorrelation (Wald and Wolfowitz), or against treatment effect in two-way ANOVA (Friedman)

● it is quite significant, in that respect, that the ANOVA(2) problem was solved well before the much simpler ANOVA(1) or two-sample location ones, and that autocorrelation problems before the location ones ...

... the storming of the Gaussian Bastille?

Wilcoxon 1945 is breaking with that spirit

● is not comparing two or several rankings anymore, as it only involves one global ranking

● addresses the most basic problems of classical statistical inference: one- and two-sample testing of location

● challenges no less than the most sacred cows of Gaussian inference: the one- and two-sample t tests

● in this respect, Wilcoxon (1945) can be considered as the Bastille Day of the nonparametric Revolution!

... after Wilcoxon ...

Mathematically, Wilcoxon's contribution was not particularly deep; and he certainly was not aware of having stormed any Bastille, but he clearly triggered an explosive development of nonparametrics:

- 58 titles in Scheffé (1943)'s early review of the subject (Wilcoxon - 2)

- 999 references in Savage's 1953 bibliography (Wilcoxon + 8)

- 3,000 references in the 1962 update (Wilcoxon + 17)

- countless ones (scholar.Google provides no less than 435,000 entries for "ranks") nowadays (Wilcoxon + 61)

The objective of this talk is a fast and unavoidably biased guided tour of that development :

Outline

1. Ranks: from distribution-freeness to group invariance
    1.1. Ranks
    1.2. Hodges-Lehmann and Chernoff-Savage
    1.3. Group invariance
2. Efficiency: from parametric to semiparametric
3. Ranks: reconciling the irreconcilable?
4. Conclusions

Outline

1. Ranks: from distribution-freeness to group invariance

    1.1. Ranks

    1.2. Hodges-Lehmann and Chernoff-Savage

    1.3. Group invariance

2. Efficiency: from parametric to semiparametric

3. Ranks: reconciling the irreconcilable?

4. Conclusions

1.1 Ranks

- Observations: $\mathbf{X} := (X_1,\ X_2,\ \ldots,\ X_n)$
- Ordered observations (order statistic):

$$\mathbf{X}_{(\ )} := X_{\min} := X_{(1)} \leq X_{(2)}, \leq \ldots \leq X_{(n)} =: X_{\max}$$

- Ranks: R_i such that $X_{(R_i)} = X_i$

or $R_i := \#\{j \mid X_j \leq X_i\}$

- Vector of ranks: $\mathbf{R}^{(n)} := (R_1,\ R_2,\ \ldots,\ R_n)$

provided that $\mathbf{X}$ is continuous, with probability one, a (random) permutation of $(1,\ 2,\ \ldots,\ n)$

If moreover the X_i 's are i.i.d., with some unspecified density over $\mathbb{R}$: then, the distribution of $\mathbf{R}^{(n)}$ is uniform over the $n!$ permutations of $(1, ..., n)$

- Advantage of $\mathbf{R}^{(n)}$ over $\mathbf{X}^{(n)}$: unlike $\mathbf{X}^{(n)}$, the distribution of which is unknown, the distribution of $\mathbf{R}^{(n)}$ is known , allowing for exact inference, robust to misspecification of f (e.g., Gaussian assumptions)
- Disadvantage of $\mathbf{R}^{(n)}$ with respect to $\mathbf{X}^{(n)}$: loss of information = loss of efficiency! (?)
- How large is that loss of information?
- The price for robustness? Conflicting objectives: robustness and efficiency

the two-sample Wilcoxon test

a closer look at, e.g., Wilcoxon's two sample (rank sum) test

- null hypothesis $\mathcal{H}_0$: $X_1, \ldots, X_m, X_{m+1}, \ldots, X_n$ iid, with unspecified density
- alternative $\mathcal{H}_1$: $X_1, \ldots, X_m, X_{m+1} - \theta, \ldots, X_n - \theta$ iid, unspecified density, for some $\theta > 0$
- test statistic: $S_W^{(n)} := \sum_{i=m+1}^{n} R_i$ distribution-free under $\mathcal{H}_0$
- Reject $\mathcal{H}_0$ for "large values" of $S_W^{(n)}$
- unlike the Student test, does not require Gaussian f

Outline

1. Ranks: from distribution-freeness to group invariance
    1.1. Ranks
    1.2. Hodges-Lehmann and Chernoff-Savage
    1.3. Group invariance
2. Efficiency: from parametric to semiparametric
3. Ranks: reconciling the irreconcilable?
4. Conclusions

A golden age?

Wilcoxon (1945) soon was followed by an explosion of rank-based methods :

● van der Waerden (1952, 1953); Fraser (1957): Location ● Mood (1950); Brown and Mood (1950); Kruskal (1951); Kruskal and Wallis (1952): One-way analysis of variance ● Mood (1950); Brown and Mood (1950); Benard and van Elteren (1953); Terpstra (1955-56) : Random blocks ● Hoeffding (1950), Terry (1952): Regression ● Savage (1956); Ansari and Bradley (1960); Capon (1961); Klotz (1962): Scale ● Wald and Wolfowitz (1944); Hoeffding (1948, 1951); Noether (1949); Dwass (1953, 1955); Fraser (1956); Motoo (1957), and many others on asymptotic normality ● ...

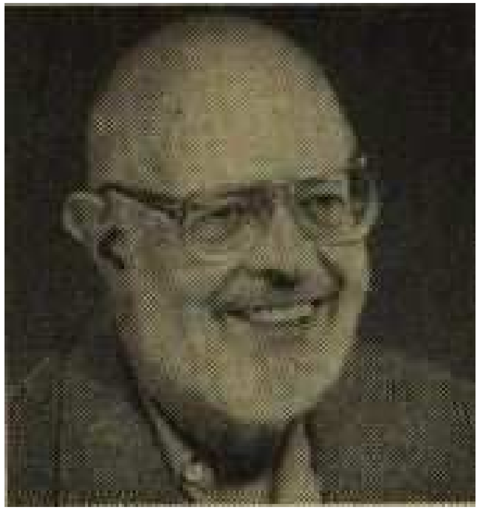
These methods were aiming mainly at robustness against departures of the Gaussian assumption and computational simplicity—they were heuristic, ad hoc, and piecemeal procedures, not expected to be particularly powerful

Pitman and Noether, ...

... empirical evidence of the power of all these methods soon emerged from practice, and, much to general amazement, could be confirmed by the newly introduced concept of *Asymptotic Relative Efficiency* (ARE) introduced by Pitman (1948, unpublished) and popularized by Noether (1950, 1955) ...

Hodges-Lehmann

... but the amazement culminated in 1956 (Wilcoxon + 11), with a most striking result by Hodges and Lehmann



Hodges and Lehmann (1956). The efficiency of some nonparametric competitors of the t-test, *Annals of Mathematical Statistics* **27**, 324-335

Hodges-Lehmann (continued)

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

In the worst case, Wilcoxon thus only requires 13.6% more observations than Student in order to achieve comparable performances! But,

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

and the benefits of unrestricted validity are invaluable ...

van der Waerden tests

Since the Normal distribution is playing such a central role, the idea of considering, for the same location problem, a statistic of the form

$$S^{(n)}_{vdW} := \sum_{i=m+1}^{n} \Phi^{-1}\left(\frac{R_i}{n+1}\right),$$

where $u \mapsto \Phi^{-1}(u)$ denotes the standard normal quantile function, was proposed

by several authors among which the eminent Dutch algebraist B.L. van der Waerden (1952) (for simplicity, we call them *van der Waerden statistics*)

van der Waerden

indeed,

- $S^{(n)}_{vdW}$ is still distribution-free (being a function of ranks)
- if the actual underlying density is normal, then $S^{(n)}_{vdW} \approx$ Student statistic
- hence, at the normal, same performances (asymptotically) as Student, which is optimal at the Gaussian

Then, even more surprising perhaps than Hodges and Lehmann,

...

Chernoff-Savage

Chernoff, H., and I.R. Savage (1958). Asymptotic normality and efficiency of certain nonparametric tests, *Annals of Mathematical Statistics* **29**, 972-994.

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

an infimum which is attained at Gaussian f only!! Hence,

- van der Waerden is always strictly better than Student, except at the normal where they are equally good ...
- ... but van der Waerden is also uniformly valid, which is not the case of Student

- ... should put Student and much of everyday practice out of business!
- ... above all, raises several very fundamental questions ...

questions (in the 1960's)

- WHAT IS IT THAT MAKES RANKS THAT EFFICIENT?
- ARE RANKS THE ONLY STATISTICAL "OBJECTS" ENJOYING SUCH ATTRACTIVE DISTRIBUTION-FREENESS/EFFICIENCY PROPERTIES?

answers (today, after a long history)

- maximal invariance
- ... and its relation to semiparametric efficiency

Too early (in the 1960's) for the idea of semiparametric efficiency; but the idea of invariance and invariant tests is present from the beginning (e.g. in Hotelling and Pabst 1936)

Outline

1. Ranks: from distribution-freeness to group invariance

    1.1. Ranks

    1.2. Hodges-Lehmann and Chernoff-Savage

    1.3. Group invariance

2. Efficiency: from parametric to semiparametric

3. Ranks: reconciling the irreconcilable?

4. Conclusions

1.3. Group invariance

Assume that $\mathbf{X} := (X_1,\ X_2,\ \ldots,\ X_n)$ are iid, with unspecified nonvanishing density f in the class $\mathcal{F}$ of all nonvanishing densities over $\mathbb{R}$ (independent white noise) . Let $\mathcal{P}^{(n)} = \left\{ \mathrm{P}^{(n)}_f \mid f \in \mathcal{F} \right\}$.

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

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

mapping $(x_1, \ldots, x_n) \in \mathbb{R}$ onto $g_h(x_1, \ldots, x_n) := (h(x_1), \ldots, h(x_n)) \in \mathbb{R}$

Then,

- $\mathcal{G}$ is a generating group for $\mathcal{P}^{(n)}$, in the sense that for all $\mathrm{P}^{(n)}_{f_1}$, $\mathrm{P}^{(n)}_{f_2}$ in $\mathcal{P}^{(n)}$, there exists $g_h \in \mathcal{G}$ such that $(X_1, \ldots, X_n) \sim \mathrm{P}^{(n)}_{f_1}$ iff $g_h(X_1, \ldots, X_n) \sim \mathrm{P}^{(n)}_{f_2}$
- the vector of ranks $\mathbf{R}^{(n)}$ is maximal invariant for $\mathcal{G}$, 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.

Other "ranks"

- $\mathbf{X} := (X_1,\ X_2,\ \ldots,\ X_n)$ iid, with unspecified nonvanishing symmetric (w. r. t. 0) density f in the class $\mathcal{F}_+$ of all nonvanishing densities over $\mathbb{R}$ (independent symmetric white noise). Let $\mathcal{P}^{(n)} = \left\{\mathrm{P}_f^{(n)} \mid f \in \mathcal{F}_+\right\}$: maximal invariant $=$ the signs and the ranks of absolute values ("signed ranks")
- $\mathbf{X} := (X_1,\ X_2,\ \ldots,\ X_n)$ iid, 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 $\mathcal{P}^{(n)} = \left\{\mathrm{P}_f^{(n)} \mid f \in \mathcal{F}_0\right\}$: maximal invariant $=$ the signs and the ranks (Hallin, Vermandele, and Werker, *Annals of Statistics* 2006)

Other “ranks”

- $\mathbf{X} := (X_1,\ X_2,\ \ldots,\ X_n)$ 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 $\mathcal{P}^{(n)} = \left\{ \mathrm{P}_f^{(n)} \mid f \in \mathcal{F}_0 \right\}$: maximal invariant = the signs

(Elliptical models)

- $\mathbf{X} := (\mathbf{X}_1,\ \mathbf{X}_2,\ \ldots,\ \mathbf{X}_n)$ iid, with elliptical density

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

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 $\mathcal{P}^{(n)} = \left\{\mathrm{P}^{(n)}_{\boldsymbol{\theta};f_1} \mid f_1 \in \mathcal{F}^+\right\}$,
where $\mathcal{F}^+$ is the class of all standardized nonvanishing densities over $\mathbb{R}^+$:

maximal invariant = the unit vectors
$\mathbf{U}_i := \mathbf{V}^{-1/2}(\mathbf{X}_i - \boldsymbol{\mu})\ /\ \|(\mathbf{X}_i - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{X}_i - \boldsymbol{\mu})\|$ and the ranks R_i of the "distances" $\|(\mathbf{X}_i - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{X}_i - \boldsymbol{\mu})\|$

Multivariate signs $\mathbf{U}_i$ and ranks R_i (Hallin and Paindaveine, *Annals of Statistics* 2002)

(Independent Component Analysis)

- $\mathbf{X} := (\mathbf{X}_1,\ \mathbf{X}_2,\ \ldots,\ \mathbf{X}_n)$ iid, where

$$\mathbf{X}_i = M\mathbf{Z}_i \quad Z_{i1}, \ldots Z_{ik} \text{ independent,}$$

with median 0 and otherwise unspecified distinct densities:

Generating group: marginal continuous order- and origin-preserving transformations

Maximal invariant $=$ the marginal ranks and signs of the $\mathbf{Z}_i$'s (Hallin, Oja, and Paindaveine, 2006)

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

...

... what do we mean exactly with efficiency and semiparametric efficiency?

Let us start with (parametric) efficiency ...

Outline

1. Ranks: from distribution-freeness to group invariance

2. Efficiency: from parametric to semiparametric

    2.1. Parametric efficiency

    2.2. Hájek projections

    2.3. Ranks in time series models

    2.4. Parametric efficiency in the presence of nuisance

    2.5. Semiparametric efficiency

3. Ranks: reconciling the irreconcilable?

4. Conclusions

Outline

1. Ranks: from distribution-freeness to group invariance

2. Efficiency: from parametric to semiparametric

    2.1. Parametric efficiency

    2.2. Hájek projections

    2.3. Ranks in time series models

    2.4. Parametric efficiency in the presence of nuisance

    2.5. Semiparametric efficiency

3. Ranks: reconciling the irreconcilable?

4. Conclusions

2.1. Parametric optimality

Throughout, we consider semiparametric models, namely, models under which the distribution of some observation $\mathbf{X} := (X_1,\ X_2,\ \ldots,\ X_n)$ belongs to a family of the form

$$\mathcal{P}^{(n)} = \left\{ \mathrm{P}^{(n)}_{f;\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \boldsymbol{\Theta},\ f \in \mathcal{F} \right\}$$

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. Assume that

- the fixed-f parametric submodels $\mathcal{P}^{(n)}_{f} := \left\{ \mathrm{P}^{(n)}_{f;\boldsymbol{\theta}} \mid \boldsymbol{\theta} \in \boldsymbol{\Theta} \right\}$ are LAN (see below)
- the fixed-$\boldsymbol{\theta}$ nonparametric submodels $\mathcal{P}^{(n)}_{\boldsymbol{\theta}} := \left\{ \mathrm{P}^{(n)}_{f;\boldsymbol{\theta}} \mid\ f \in \mathcal{F} \right\}$ are generated by some group $\mathcal{G}^{(n)}_{\boldsymbol{\theta}}$ with maximal invariant $\mathbf{R}^{(n)}(\boldsymbol{\theta})$

LAN w.r.t. $\boldsymbol{\theta}$ (at given f)

Some mathematics ...

Lucien Le Cam's LAN

("It is easy once you know ...")

Under $\mathrm{P}^{(n)}_{\boldsymbol{\theta};f}$, as $n \to \infty$,

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

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

• the random vector $\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f}$ is called the central sequence (localized at $\boldsymbol{\theta}$)

• the deterministic matrix $\boldsymbol{\Gamma}_{\boldsymbol{\theta};f}$ is called the information matrix (at $\boldsymbol{\theta}$)

parametric efficiency (at given f)

Skipping technicalities,

- under $\mathrm{P}^{(n)}_{\boldsymbol{\theta}+n^{-1/2}\boldsymbol{\tau};f}$, $\boldsymbol{\tau}\in\mathbb{R}^m$, the central sequence

  $\boxed{\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f} \text{ is asymptotically } \mathcal{N}\left(\boldsymbol{\Gamma}_{\boldsymbol{\theta};f}\boldsymbol{\tau},\boldsymbol{\Gamma}_{\boldsymbol{\theta};f}\right)}$

- 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 $\boxed{\boldsymbol{\Delta}\sim\mathcal{N}\left(\boldsymbol{\Gamma}_{\boldsymbol{\theta};f}\boldsymbol{\tau},\boldsymbol{\Gamma}_{\boldsymbol{\theta};f}\right), \quad \boldsymbol{\tau}\in\mathbb{R}^m}$

- that is, for instance, optimal α-level tests of $\boldsymbol{\theta}=\boldsymbol{\theta}_0$ achieving at $\mathrm{P}^{(n)}_{\boldsymbol{\theta}_0+n^{-1/2}\boldsymbol{\tau};f}$ power $1-F_{m;\boldsymbol{\tau}'\boldsymbol{\Gamma}^{-1}_{\boldsymbol{\theta}_0;f}\boldsymbol{\tau}}(\chi^2_{m;1-\alpha})$, where $F_{m;\lambda}$ stands for the noncentral chi square distribution function with m degrees of freedom and noncentrality parameter λ

- or optimal estimators $\hat{\boldsymbol{\theta}}^{(n)}$ of $\boldsymbol{\theta}$ such that

$$n^{1/2}(\hat{\boldsymbol{\theta}}^{(n)}-\boldsymbol{\theta})\xrightarrow{\mathcal{L}}\boldsymbol{\Gamma}^{-1}_{\boldsymbol{\theta};f}\boldsymbol{\Delta}\sim\mathcal{N}\left(\mathbf{0},\boldsymbol{\Gamma}^{-1}_{\boldsymbol{\theta};f}\right)$$

parametric efficiency (at given f)

- Moreover, optimality is achieved by treating the central sequence $\mathbf{\Delta}^{(n)}_{\boldsymbol{\theta};f}$ exactly as one would the observation $\mathbf{\Delta}$ in the limit Gaussian shift model
- that is, for instance, by means of optimal tests statistics for $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ of the form $Q_F := (\mathbf{\Delta}^{(n)}_{\boldsymbol{\theta}_0;f})'\mathbf{\Gamma}^{-1}_{\boldsymbol{\theta}_0;f}\mathbf{\Delta}^{(n)}_{\boldsymbol{\theta}_0;f} \approx \chi^2_m$
- or optimal estimators $\hat{\boldsymbol{\theta}}^{(n)}$ (of the one-step form) such that

$$n^{1/2}(\hat{\boldsymbol{\theta}}^{(n)} - \boldsymbol{\theta}) = \mathbf{\Gamma}^{-1}_{\boldsymbol{\theta};f}\mathbf{\Delta}(\boldsymbol{\theta};f) + o_{\mathrm{P}}(1) \approx \mathcal{N}\left(\mathbf{0}, \mathbf{\Gamma}^{-1}_{\boldsymbol{\theta};f}\right)$$

parametric efficiency (at given f and $\boldsymbol{\theta}$) is characterized by the Gaussian shift model

$$\boldsymbol{\Delta} \sim \mathcal{N}\left(\boldsymbol{\Gamma}_{\boldsymbol{\theta};f}\boldsymbol{\tau}, \boldsymbol{\Gamma}_{\boldsymbol{\theta};f}\right), \quad \boldsymbol{\tau} \in \mathbb{R}^m$$

hence by the information matrix $\boldsymbol{\Gamma}_{\boldsymbol{\theta};f}$

Outline

1. Ranks: from distribution-freeness to group invariance

2. Efficiency: from parametric to semiparametric

   2.1. Parametric efficiency

   2.2. Hájek projections

   2.3. Ranks in time series models

   2.4. Parametric efficiency in the presence of nuisance

   2.5. Semiparametric efficiency

3. Ranks: reconciling the irreconcilable?

4. Conclusions

then came Jaroslav Hájek ...

It was high time, in the mid-sixties, for some unification of what remained a scattered and unstructured collection of heuristic methods. An invaluable step in that direction was accomplished by a Czech mathematician: Jaroslav Hájek

Hájek projections

Hájek mainly considers linear models (regression models). Although he does not state it under that form (as the Le Cam theory at that time was not fully available), what he proposes is the construction of rank-based central sequences resulting from projecting the "parametric" $\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f}$ (associated with some reference density f) onto the ranks $\mathbf{R}^{(n)}(\boldsymbol{\theta})$ of residuals:

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

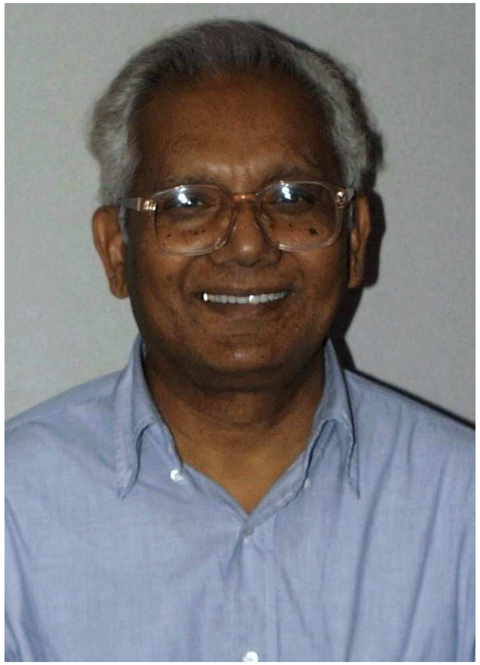
and, under very general conditions, he actually shows (a corollary to his asymptotic representation theorem) that, under $\mathrm{P}^{(n)}_{\boldsymbol{\theta};f}$, in the linear model with i.i.d. errors,

$$\underset{\sim}{\boldsymbol{\Delta}}^{(n)}_{\boldsymbol{\theta};f} = \boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f} + o_{\mathrm{P}}(1)$$

(this, however, is a modern, ex-post account of Hájek's approach)

the Hájek era (1967-1985)

- Hájek's ideas were summarized in a remarkable 1967 monograph : Hájek and Šidák (1967), *Theory of Rank Tests*
- these ideas (after Hájek's untimely death) were systematized in several subsequent monographs, of which Puri and Sen (1985), *Nonparametric Methods in General Linear Models* is perhaps the most representative



the twilight of ranks?

In 1985, the theory of rank-based inference seemed pretty complete and, everybody felt that, due to its permutational nature, it was inherently limited to linear models with independent errors

Under its “conditioning of central sequences” form, however, the Hájek methodology potentially applies in the much more general context of LAN models

Outline

1. Ranks: from distribution-freeness to group invariance

2. Efficiency: from parametric to semiparametric

    2.1. Parametric efficiency

    2.2. Hájek projections

    2.3. Ranks in time series models

    2.4. Parametric efficiency in the presence of nuisance

    2.5. Semiparametric efficiency

3. Ranks: reconciling the irreconcilable?

4. Conclusions

ranks and time series: a catalytic effect

- although attention had been focused on applications of ranks in linear models, some of the earliest rank-based methods actually were time-series methods: tests based on runs, the turning point test, Wald and Wolfowitz (1943) ... actually are dealing with problems of serial dependence
- taking advantage of recent LAN results for ARMA models and extending to a serial context the Hájek projection idea, Hallin, Ingenbleek and Puri (*Annals of Statistics* 1985, 1988) obtain parametrically efficient rank-based methods for ARMA time series based on a new class of *linear serial rank statistics*
- with a new asymptotic representation theorem involving an additional term

the irruption of nonadaptivity

- in the ARMA case, that additional term disappears, and $\underset{\sim}{\boldsymbol{\Delta}}^{(n)}_{\boldsymbol{\theta};f} = \boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f} + o_{\mathrm{P}}(1)$ as in linear models with iid errors: a rank-based reconstruction of $\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f}$ is thus possible, and ranks achieve parametric efficiency
- later on, exploiting a similar methodology, Benghabrit and Hallin (1992, 1996), and Akharif and Hallin (*Annals of Statistics* 2003) derive some rank-based methods for problems involving nonlinear time series (bilinear, random coefficient AR, ...)
- it appears however that, contrary to the case of linear models with independent errors, and contrary to ARMA models, the additional term in the asymptotic representation result does not disappear: a rank-based reconstruction of $\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f}$ is thus impossible, and ranks do not achieve parametric efficiency anymore

a "nasty, ugly little fact"?

Why is that so? how does that come ? is this "the tragedy of a beautiful theory, killed by a nasty, ugly little fact [a]"?

Not quite so! Actually, that nasty additional term is nothing but the consequence of non-adaptivity in the semiparametric sense ...

[a] Thomas H. Huxley, cited in Stigler (2006), "The tragic story of maximum likelihood"

Outline

1. Ranks: from distribution-freeness to group invariance

2. Efficiency: from parametric to semiparametric

    2.1. Parametric efficiency

    2.2. Hájek projections

    2.3. Ranks in time series models

    2.4. Parametric efficiency in the presence of nuisance

    2.5. Semiparametric efficiency

3. Ranks: reconciling the irreconcilable?

4. Conclusions

2.2. Parametric efficiency in the presence of a nuisance: a parenthesis

Assume that $\boldsymbol{\theta} = (\theta_1, \theta_2)$, and that inference is to be made about θ_1, while θ_2 is a nuisance; the central sequence $\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f}$ similarly decomposes into $(\Delta^{(n)}_{\boldsymbol{\theta};f;1}, \Delta^{(n)}_{\boldsymbol{\theta};f;2})$, and the information matrix into

$$\boldsymbol{\Gamma}_{\boldsymbol{\theta};f} = \begin{pmatrix} \Gamma_{\boldsymbol{\theta};f;11} & \Gamma_{\boldsymbol{\theta};f;12} \\ \Gamma_{\boldsymbol{\theta};f;12} & \Gamma_{\boldsymbol{\theta};f;22} \end{pmatrix}$$

Referring to the limit Gaussian shift, it is easy to understand that such inference on θ_1 should be based on

$$\Delta^{(n)}_{\boldsymbol{\theta};f;1} - \Gamma_{\boldsymbol{\theta};f;12}\Gamma^{-1}_{\boldsymbol{\theta};f;22}\Delta^{(n)}_{\boldsymbol{\theta};f;2},$$

the residual of the regression of $\Delta^{(n)}_{\boldsymbol{\theta};f;1}$ on $\Delta^{(n)}_{\boldsymbol{\theta};f;2}$ in the covariance $\boldsymbol{\Gamma}_{\boldsymbol{\theta};f}$, or the $\boldsymbol{\Gamma}_{\boldsymbol{\theta};f}$-projection of the θ_1-central sequence orthogonal to the space of the θ_2-central sequence

- indeed, a local perturbation $n^{-1/2}\tau_2$ of θ_2 induces (via Le Cam's Third Lemma) on the asymptotic distribution of $(\Delta^{(n)}_{\boldsymbol{\theta};f;1}, \Delta^{(n)}_{\boldsymbol{\theta};f;2})$ a shift $(\mathbf{\Gamma}_{\boldsymbol{\theta};f;12}\;\tau_2, \mathbf{\Gamma}_{\boldsymbol{\theta};f;22}\;\tau_2)$;
the resulting shift for the residual $\Delta^{(n)}_{\boldsymbol{\theta};f;1} - \mathbf{\Gamma}_{\boldsymbol{\theta};f;12}\mathbf{\Gamma}^{-1}_{\boldsymbol{\theta};f;22}\Delta^{(n)}_{\boldsymbol{\theta};f;2}$, is thus $\mathbf{\Gamma}_{\boldsymbol{\theta};f;12}\;\tau_2 - \mathbf{\Gamma}_{\boldsymbol{\theta};f;12}\mathbf{\Gamma}^{-1}_{\boldsymbol{\theta};f;22}\mathbf{\Gamma}_{\boldsymbol{\theta};f;22}\;\tau_2 = 0$:

- this residual therefore is insensitive to local perturbations of θ_2

parametric efficiency (at given f and $\boldsymbol{\theta}$) for θ_1 when θ_2 is a nuisance is characterized by the Gaussian shift model

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

hence by the information matrix $\Gamma_{\boldsymbol{\theta};f;11} - \Gamma_{\boldsymbol{\theta};f;12}\Gamma^{-1}_{\boldsymbol{\theta};f;22}\Gamma'_{\boldsymbol{\theta};f;12}$

Outline

1. Ranks: from distribution-freeness to group invariance

2. Efficiency: from parametric to semiparametric

    2.1. Parametric efficiency

    2.2. Hájek projections

    2.3. Ranks in time series models

    2.4. Parametric efficiency in the presence of nuisance

    2.5. Semiparametric efficiency

3. Ranks: reconciling the irreconcilable?

4. Conclusions

2.5. Semiparametric efficiency

Previous conclusions on parametric efficiency are valid under correctly specified density f only!!

In practice, f is the nuisance

- this nonparametric nuisance is treated in the same way as the parametric nuisance θ_2: projection of the central sequence along the space generated by the shifts induced by the variations of densities in the “vicinity” of f: the “tangent space“, characterized by the “least favorable” parametric perturbation of f
- Classical reference: Bickel, Klaassen, Ritov, and Wellner (1993)

2.5. Semiparametric efficiency

- this yields a projected central sequence , or "semiparametrically efficient (at f) central sequence" $\mathbf{\Delta}^{(n)*}_{\boldsymbol{\theta};f}$, with (asymptotic) covariance $\mathbf{\Gamma}^{*}_{\boldsymbol{\theta};f} \leq \mathbf{\Gamma}_{\boldsymbol{\theta};f}$—the "semiparametrically efficient (at f) information matrix"

- whenever $\mathbf{\Gamma}^{*}_{\boldsymbol{\theta};f} = \mathbf{\Gamma}_{\boldsymbol{\theta};f}$: the model is "adaptive" at f
- in general, $\mathbf{\Gamma}^{*}_{\boldsymbol{\theta};f} < \mathbf{\Gamma}_{\boldsymbol{\theta};f}$: the cost of not knowing the "true" density, at f, is strictly positive

semiparametric efficiency (at given f and $\boldsymbol{\theta}$) is characterized by the Gaussian shift model

$$\boldsymbol{\Delta}^* \sim \mathcal{N}\left(\boldsymbol{\Gamma}^*_{\boldsymbol{\theta};f}\boldsymbol{\tau}, \boldsymbol{\Gamma}^*_{\boldsymbol{\theta};f}\right), \ \tau \in \mathbb{R}^m$$

hence by the semiparametrically efficient (at f) information matrix $\boldsymbol{\Gamma}^*_{\boldsymbol{\theta};f}$

tangent spaces

- projections along tangent spaces typically are not easily computed
- ... more serious: the semiparametrically efficient (at f) central sequence $\boldsymbol{\Delta}^{(n)*}_{\boldsymbol{\theta};f}$ is asymptotically $\mathcal{N}\left(\mathbf{0}, \boldsymbol{\Gamma}^{*}_{\boldsymbol{\theta};f}\right)$ under density f only (indeed the projection along the tangent space at f depends on f)

hence, inference based on $\boldsymbol{\Delta}^{(n)*}_{\boldsymbol{\theta};f}$ is valid under density f only

... the semiparametrically efficient (at f) central sequence $\mathbf{\Delta}^{(n)*}_{\boldsymbol{\theta};f}$ thus looks like a useless tool ...

... unless an estimated density $\hat{f}^{(n)}$can be obtained such that $\mathbf{\Delta}^{(n)*}_{\boldsymbol{\theta};\hat{f}^{(n)}} - \mathbf{\Delta}^{(n)*}_{\boldsymbol{\theta};f} = o_{\mathrm{P}}(1)$ at "all" f:

kernel density estimators, additional regularity assumptions, slow

convergence rates, sample splitting and other niceties ...

Outline

1. Ranks: from distribution-freeness to group invariance

2. Efficiency: from parametric to semiparametric

3. Ranks: reconciling the irreconcilable?

    3.1. From tangent space to Hájek projection

    3.2. Example: rank-based inference for nonlinear time series

    3.3. Example: rank-based inference for shape

4. Conclusions

Outline

1. Ranks: from distribution-freeness to group invariance

2. Efficiency: from parametric to semiparametric

3. Ranks: reconciling the irreconcilable?

    3.1. From tangent space to Hájek projection

    3.2. Example: rank-based inference for nonlinear time series

    3.3. Example: rank-based inference for shape

4. Conclusions

The statistical principle behind Hájek's approach is invariance.

The Invariance Principlestipulates that "when a statistical problem is invariant under the action of some group of transformations, one should restrict to invariant statistical procedures"

it has been assumed that the fixed-$\boldsymbol{\theta}$ submodels of our semiparametric models are invariant w.r.t. groups $\mathcal{G}_{\boldsymbol{\theta}}$, with maximal invariant $\mathbf{R}^{(n)}(\boldsymbol{\theta})$ (typically, the ranks of some $\boldsymbol{\theta}$-residuals)

the invariant statistics (in those models) thus are the functions of $\mathbf{R}^{(n)}(\boldsymbol{\theta})$

from tangent space to Hájek projections

a natural idea thus consists in considering the invariant statistics which are closest to the central sequences by projecting $\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f}$ onto the σ-field generated by $\mathbf{R}^{(n)}(\boldsymbol{\theta})$, yielding (up to $o_{\mathrm{P}}(1)$'s)

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

... which is what Hájek actually is doing (without telling—the concept of a central sequence was not available to him)

- being $\mathbf{R}^{(n)}(\boldsymbol{\theta})$-measurable, $\underset{\sim}{\mathbf{\Delta}}^{(n)}_{\boldsymbol{\theta};f}$ is a distribution-free statistic (in the the fixed-$\boldsymbol{\theta}$ submodel)
- in linear models with i.i.d. errors, or in ARMA models with i.i.d. innovations, this yields a rank-based version of the central sequence : $\underset{\sim}{\mathbf{\Delta}}_{\boldsymbol{\theta};f} = \mathbf{\Delta}^{(n)}_{\boldsymbol{\theta};f} + o_{\mathrm{P}}(1)$; these models are adaptive
- this does not hold anymore in nonlinear time series models, which are NOT adaptive

Hájek projections and tangent spaces

- in 2002, Hallin and Werker (*Bernoulli* 2002) under very general conditions show that, actually, under $\mathrm{P}^{(n)}_{\boldsymbol{\theta};f}$,

$$\underset{\sim}{\boldsymbol{\Delta}}^{(n)}_{\boldsymbol{\theta};f} = \boldsymbol{\Delta}^{(n)*}_{\boldsymbol{\theta};f} + o_{\mathrm{P}}(1)$$

- It appears that the "nasty additional term" in the 1985 Hallin-Ingenbleek-Puri asymptotic representation, actually, is (asymptotically) the projection $\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f} - \boldsymbol{\Delta}^{(n)*}_{\boldsymbol{\theta};f}$ of $\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f}$ onto the tangent space associated with unspecified densities

Hájek projections and tangent spaces

it follows that

- $\underset{\sim}{\boldsymbol{\Delta}}^{(n)}_{\boldsymbol{\theta};f}$ is an invariant (rank-based) distribution-free version of the semiparametrically efficient (at f) central sequence;
- contrary to the classical version of the latter, its distribution does not depend on the underlying density, and it thus allows for valid (and even distribution-free) semiparametrically efficient-at-f inference on $\boldsymbol{\theta}$
- Hájek projections are doing the same job as tangent space projections, without requiring the (often nontrivial) computation of the latter: *the ranks do it for you!*, and with the (invaluable) additional advantages of distribution-freeness

- $\mathrm{E}\left[\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f} \mid \mathbf{R}^{(n)}(\boldsymbol{\theta})\right]$ is the “exact score version” of $\underset{\sim}{\boldsymbol{\Delta}}_{\boldsymbol{\theta};f}$; 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 $\underset{\sim}{\boldsymbol{\Delta}}_{\boldsymbol{\theta};\hat{f}^{(n)}}$, where $\hat{f}^{(n)}$ is some appropriate density estimator
- without 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, we also split the sample, into two mutually independent parts: the invariant and distribution-free part on one hand (the ranks), the “order statistic” 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

Outline

1. Ranks: from distribution-freeness to group invariance

2. Efficiency: from parametric to semiparametric

3. Ranks: reconciling the irreconcilable?

    3.1. From tangent space to Hájek projection

    3.2. Example: rank-based inference for nonlinear time series

    3.3. Example: rank-based inference for shape

4. Conclusions

Outline

1. Ranks: from distribution-freeness to group invariance

2. Efficiency: from parametric to semiparametric

3. Ranks: reconciling the irreconcilable?

    3.1. From tangent space to Hájek projection

    3.2. Example: rank-based inference for nonlinear time series

    3.3. Example: rank-based inference for shape

4. Conclusions

3.3. Rank-based inference for shape

Writing $\mathbf{X} \sim \mathrm{P}^{(n)}_{\boldsymbol{\theta};f_1}$, $\boldsymbol{\theta} = (\boldsymbol{\mu}, \sigma, \mathbf{V})$ and $\mathcal{P}^{(n)} = \left\{\mathrm{P}^{(n)}_{\boldsymbol{\theta};f_1} \mid f_1 \in \mathcal{F}^+\right\}$, where $\mathcal{F}^+$ is the class of all standardized nonvanishing densities over $\mathbb{R}^+$ when

- $\mathbf{X} := (\mathbf{X}_1,\ \mathbf{X}_2,\ \ldots,\ \mathbf{X}_n)$ iid, with elliptical density

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

over $\mathbb{R}^k$ (independent elliptical white noise with location $\boldsymbol{\mu}$, shape $\mathbf{V}$, scale σ, and standardized radial density f_1), maximal invariant are

- the ranks R_i of the "radial distances"

$$d_i := \|(\mathbf{X}_i - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{X}_i - \boldsymbol{\mu})\|, \text{ and}$$

- the unit vectors $\mathbf{U}_i := (\mathbf{X}_i - \boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{X}_i - \boldsymbol{\mu})\ /\ d_i$
- Multivariate signs $\mathbf{U}_i$ and
- ranks R_i

central sequences for shape (at f): from parametric to semiparametric and rank-based (Hallin and Paindaveine, *Annals of Statistics*, 2006)

- parametric

$$\boldsymbol{\Delta}^{(n)}_{\boldsymbol{\theta};f_1} := \frac{1}{2} n^{-1/2} \mathbf{M}_k \left(\mathbf{V}^{\otimes 2}\right)^{-1/2} \sum_{i=1}^{n} \varphi_{f_1}\left(\frac{d_i}{\sigma}\right) \frac{d_i}{\sigma} \mathrm{vec}\left(\mathbf{U}_i \mathbf{U}_i' - \mathbf{I}_k\right)$$

- semiparametrically efficient

$$\boldsymbol{\Delta}^{(n)*}_{\boldsymbol{\theta};f_1} = \frac{1}{2} n^{-1/2} \mathbf{M}_k \left[\mathbf{I}_{k^2} - \frac{1}{k}\mathbf{J}_k\right] \sum_{i=1}^{n} \varphi_{f_1}\left(\frac{d_i}{\sigma}\right) \frac{d_i}{\sigma} \mathrm{vec}\left(\mathbf{U}_i \mathbf{U}_i'\right)$$

- rank-based

$$\underset{\sim}{\boldsymbol{\Delta}}^{(n)}_{\boldsymbol{\theta};f_1} := \frac{1}{2} n^{-1/2} \mathbf{M}_k \left[\mathbf{I}_{k^2} - \frac{1}{k}\mathbf{J}_k\right] \sum_{i=1}^{n} K_{f_1}\left(\frac{R_i}{n+1}\right) \mathrm{vec}\left(\mathbf{U}_i \mathbf{U}_i'\right),$$

with $K_{f_1}(u) := (\varphi_{f_1} \circ F_1)^{-1}(u)\, F_1^{-1}(u)$

Outline

1. Ranks: from distribution-freeness to group invariance
2. Efficiency: from parametric to semiparametric
3. Ranks: reconciling the irreconcilable?
4. Conclusions

4. Conclusions

- rank-based methods (more generally, the “maximal invariant” ones) are flexible, and apply in a very broad class of statistical models, much beyond the traditional context of linear models with independent observations
- rank-based procedures are powerful—often, making Gaussian or pseudo-Gaussian methods non-admissible—with the additional benefits of distribution-freeness
- rank-based procedures (more generally, the “maximal invariant” ones) are “optimal” (achieving semiparametric efficiency—the best we can hope for in presence of unspecified densities) although they are simpler and more robust (distribution-freeness) than “classical” semiparametric procedures

- 63 years after Wilcoxon's pioneering paper, the "quick and easy tricks" have grown into a full body of efficient and modern methods

- the enemy brothers of statistics, efficiency (semiparametric) and robustness (distribution-freeness = 100% resistance against misspecified densities), can be reconciled!

THANK YOU, AND A HAPPY BIRTHDAY, MR WILCOXON!