Marc Hallin’s research activities are covering a broad spectrum of fundamental and applied statistical topics, including statistical decision, time series, random fields, density estimation, multivariate analysis, panel data, inequalities, high-dimensional and “big data” problems, quantile regression, spectral analysis, data depth, the asymptotic theory of statistical experiments, statistical applications of measure transportation, and econometrics.

In all those topics, Marc has been promoting nonparametric and semiparametric approaches. Two major threads of activities are emerging from a list of more than 220 publications:

(i) rank-based and quantile-oriented inference and

(ii) the analysis of high-dimensional time-series data.

1 Rank-based and quantile-oriented inference.

The main part of Marc’s research activity has been devoted to a decision-theoretical approach to rank-based and quantile-oriented inference, and extensions thereof. Marc’s interest in ranks goes back to the mid-eighties. After about half a century of intensive development and the seminal contribution of Hájek, as summarized in Hájek and Šidák (1967) and systematized in such monographs as Puri and Sen (1985), rank-based inference in the eighties was considered an essentially complete theory, and research activity in the area slowed down quite significantly. The theory at that point, however, was covering a somewhat narrow range of statistical models,

- being essentially limited to the context of general linear models with independent (exchangeable) observations (location, scale, regression, ANOVA, etc.),

- restricted to the analysis of univariate observations (except for methods based on componentwise rankings, which are unsatisfactory on many counts), and

- making limited use of Le Cam’s theory of statistical experiments (essentially, only Le Cam’s third Lemma was used, mainly in order to compute local powers—a practice that goes back to Hájek and Šidák 1967); as for the Bickel, Klaassen, Ritov and Wellner (1993) approach to semiparametric inference, it was not available yet.

Marc quite successfully succeeded in removing most of those restrictions, extending the scope of rank-based methods into a variety of directions:
to linear (ARMA) time series models first (the first significant steps in that direction being [32] and [48]), and, more recently ([108], [203], [213]), to semiparametric versions of the sophisticated nonlinear dynamic models considered in econometrics—discretely observed diffusions with jumps, AR-ARCH, AR-LARCH, Cox-Ingersoll-Ross processes, duration models, etc.,

to more general concepts of ranks (related to appropriate group invariance or maximal ancillary arguments), such as signs and ranks for median- or quantile-restricted models ([139] and [151]), ranks and some adequate indicators (e.g., in the context of Ornstein-Uhlenbeck processes, see [108], or for unrestrictedly heteroskedastic time series, see [96]), and

to multivariate or multiple-output settings (pseudo-Mahalanobis ranks and signs, hyperplane-based ranks and signs, and, recently, Monge-Kantorovich (center-outward) quantiles, ranks, and signs, based on measure transportation ideas: see, for instance [121], [122], [132], [140], [141], ..., [199], [216], [217], [218], [219], [226], [227], [228]).

by fully exploiting the power of Le Cam’s theory of locally asymptotically normal experiments, and the related theory of semiparametric inference (Bickel et al. 1993).

1.1 Rank-based methods and time-series analysis

Marc can be credited for a systematic introduction of rank-based methods in time-series analysis. Thanks to his contributions, rank tests now are available for

- ARMA processes ([48], [49], [102]; with a linear trend [69]), bilinear ([64], [78], [95]) and random coefficient AR models ([125]), but also fractional differentiation and long-memory ([164]),

- the sophisticated nonlinear time-series models considered in econometrics, such as Ornstein-Uhlenbeck processes ([108]), AR-ARCH and LARCH, or discretely observed diffusions and Lévy processes ([203], [213]), where Gaussian quasi-likelihoods are not always quite adequate, while the classical semiparametric tangent space approach may lead to numerically difficult problems, and, in the multivariate context,

- VARMA ([54], [74], [219], [228]) and elliptical VARMA models, possibly with trend ([122], [131], [132], [133], [135]), unit roots and cointegration ([169], [197]).

Another concept related with rank-based inference is that of (auto)regression rank scores. The concept was introduced in a regression context by Gutenbrunner and Jurečková (Annals of Statistics 1992) from a duality argument applied to Koenker and Bassett’s celebrated regression quantiles. The most attractive feature of those rank scores lies in the fact that, contrary to classical “aligned rank statistics” computed from estimated residuals, regression rank score statistics asymptotically reconstruct the actual corresponding rank-based statistics, even though exact residuals (hence exact ranks) cannot be computed from the observations due to the presence of unspecified nuisance parameters. The concept has been extended by Koul and Saleh (1995) to the time-series context, albeit in the Jaeckel style (mixing residuals and their ranks). In [103], Marc and Jana Jurečková are constructing locally asymptotically optimal tests based on such autoregression rank scores derived for linear constraints on the coefficients of an autoregressive model. The related estimation procedures are derived in [118]. In [92] and [99], the technique is applied
to the problem of autoregressive order identification, and to the problem of testing independence between two autoregressive series with unspecified coefficients. Kolmogorov-Smirnov tests based on autoregression rank scores are constructed in [116], and a very efficient method for estimating the innovation sparsity function (the inverse of the density of the unobservable innovation process at some given quantile) is proposed in [120]. Finally, [144] (joint with Jana Jurečková and Hira Koul) proposes a class of serial statistics which, contrary to Koul and Saleh’s, is entirely based on (auto)regression rank scores (thus involving multiple integrals over the quantile ranges of several lagged residuals), and asymptotically equivalent to the corresponding statistic based on the genuine, non-available residual ranks—something the Jaeckel-type statistics cannot achieve.

Generalizations of the classical Chernoff-Savage Theorem, stating that rank tests based on Gaussian scores perform uniformly better than Student tests, are established for a number of those extensions: see [72] and [105] for univariate time-series problems, [121] and [122] for non-serial and serial elliptical ones. In time series, for instance, the asymptotic relative efficiencies, with respect to daily practice correlogram-based (pseudo-Gaussian) methods, of the normal-score rank-based procedures developed in [69], are shown to be uniformly larger than one. This should be a strong incentive for bringing ranks into practice in the context.

Marc also has obtained generalizations of the no less famous Hodges and Lehmann “.864” result. In its original version, this result shows that the lower bound for the asymptotic relative efficiencies, still with respect to Student tests, of Wilcoxon-type methods for location, is .864. In a time-series context, with Student replaced by correlogram-based methods, that bound (see [105]) takes a slightly smaller .856 value. It is interesting to note that, in higher dimensions (k-dimensional elliptical observations, the Gaussian reference being Hotelling rather than Student), this Hodges-Lehmann bound, with a maximum value of .916 at dimension k = 2, is not a monotone function of k; see [121] and [122]. In the same spirit, [184] investigates several extensions of Hodges and Lehmann’s “6/π result”.

1.2 Rank-based methods and elliptical multivariate analysis without second-order moment assumptions

In a series of papers with Davy Paindaveine, Marc also showed how a rank-based approach to classical multivariate analysis problems is possible in the extended context of elliptical families (possibly, under infinite variances). They developed distribution-free yet asymptotically optimal (parametrically or semiparametrically) signed-rank methods (the ranks here are those of elliptical distances to the center, the signs are cosines between observed directions) for

- one-sample location ([121]);
- testing white noise against VARMA dependence ([122]), identifying VAR order ([133]), testing in linear models and linear models with VARMA errors ([135]);
- one- and m-sample shape problems ([140], [141], [143]) problems; homogeneity of shape and scatter ([149], [154]);
- principal components and common principal components ([161], [179], [187]).

Still in the multivariate context, [191] develops an R-estimation procedure for the mixing matrix in independent component analysis (without the assumption of symmetric component densities).
Some of these contributions are solving long-standing open problems in multivariate analysis, such as the correction ([149], [154]; see [157] for a detailed discussion) of Bartlett’s test of homogeneity of variances (here extended, under possibly infinite variances, to homogeneity of scatter) under unspecified densities—previous solutions indeed either were destroying the local power, or were losing the local maximin structure of that celebrated, daily-practice tool. Some others are introducing new asymptotic concepts, such as that of \textit{locally asymptotically curved experiments} (see [161], [179], and [187]).

1.3 Rank-based inference and semiparametric efficiency

Semiparametric models, where the underlying density or innovation density plays the role of a nuisance parameter) indeed are the general context where rank-based methods naturally come into the picture. A far-reaching result is obtained in [124], where it is shown that conditioning central sequences with respect to the maximal invariants (ranks, for instance) of appropriate generating groups yields the same results as the more traditional tangent space projections, hence leads to semiparametrically efficient inference.

That result provides a fundamental and very strong justification for considering rank-based inference, by showing that ranks (or more general invariants) actually retain all the information related with the parameter under study, while everything else (typically, an order statistic) only carries information about the nuisance (the underlying unspecified density). Ranks are thus, in a sense, implicitly rather than explicitly, performing the tangent space projections, without requiring the case-by-case derivation of least favorable directions, the explicit computation of the tangents, nor any kernel-estimation of the actual densities; they also avoid unpleasant ad hoc procedures such as sample splitting. As a corollary, ranks can reach parametric efficiency in a given model if and only if that model is adaptive in the semiparametric sense. And, of course, ranks also bring along the many advantages related to distribution-freeness: exact distributions, unbiasedness, increased robustness, ...

The above result applies to a variety of models: location, scale, regression, ANOVA, of course, but also ARMA, bilinear models (in the vicinity of the linear ones), random coefficient autoregressive models (in the vicinity of the nonrandom ones), ARCH and GARCH models, ... and, in the multivariate context, to their elliptical counterparts—including the problems listed in Section 1.2. It is also a strong argument for reviving R-estimation (see Section 1.4).

1.4 R-estimation

R-estimation is another classical topic in rank-based inference. Unlike testing, however, and despite a long history R-estimation, with the exception of linear regression, never really made its way to applications, and a widespread idea is that “ranks are fine for testing but not for estimation”.

The reason for this is twofold. Practical reasons first: in contrast with rank-based test statistics, R-estimators do not come under explicit closed forms, but as solutions of optimization procedures involving piecewise constant and nonconvex objective functions; the larger the dimension of the parameter, the trickier the computation of such estimators. Next, the asymptotic variances of R-estimators, which are needed for computing asymptotic confidence regions, typically depend on unknown \textit{cross-information quantities} of the form (in the particular case of regression)

\[
\mathcal{J}(f, g) = \int_{0}^{1} \frac{f'(F^{-1}(u))}{f(F^{-1}(u))} \frac{g'(G^{-1}(u))}{g(G^{-1}(u))} du
\]
where $F$ is the distribution function associated with the reference density $f$ used to build the scores, but $G$ and $g$ are the actual unspecified distribution function and density. Such integrals are not easily estimated. For instance, in the Wilcoxon case, for one-sample location, one has to estimate $\int g^2(z) \, dz$, where $g$ is the unspecified density of the observations. Unless $R$-estimators can be computed via some well-behaved alternative minimization (this is the case, for regression—but for regression only, of Jaeckel’s approach), $R$-estimation thus remains a theoretically attractive but practically infeasible method. This is particularly regrettable in view of the strong connexion established in Section 2.3, between rank-based methods and semiparametric efficiency.

Inspired by Le Cam’s one-step technique, Marc proposes a one-step $R$-estimator based on the same rank-based central sequence as the semiparametrically efficient rank tests described in Section 2.3. The problem with this one-step $R$-estimator is that it explicitly depends on the same unknown cross-information quantity $I(f, g)$ as above. In [141] and [162], Marc shows how an ingenious one-dimensional local likelihood maximization argument, exploiting the LAN structure of the experiment under study, provides a consistent estimation of that quantity. Under this one-step form, $R$-estimators in principle can be constructed and computed for the broad range of models considered in [124].

The method has been successfully applied to the estimation of shape matrices in elliptical families ([141]), that of regression coefficients in linear models with stable errors ([178]; unlike the much-studied OLS estimator, this one is root-$n$ consistent), the estimation of principal component and common principal components ([187]; the only available estimators of common principal components retaining consistency in the absence of Gaussian assumptions), and the estimation of mixing matrices in independent component analysis ([191]). In all those problems, the dimension of the parameter to be estimated is relatively large, and a method involving a one-dimensional optimization step only is quite welcome. The asymptotic performances of those estimators are particularly good when based on data-driven reference densities, such as skew-$t$ ones with estimated skewness and degrees of freedom accounting for the skewness and kurtosis of the actual unspecified density (which does not have to be skew-$t$). In [203] and [213], the method is used in sophisticated dynamic models from econometrics—non-Gaussian discretely observed Ornstein-Uhlenbeck processes, discretely observed diffusions with jumps, AR-ARCH, AR-LARCH, Cox-Ingersoll-Ross processes, duration models, etc.—for which Gaussian quasi-likelihood methods typically lead to poor and non-robust results or difficult implementations.

Another important application is in regression models with infinite-variance $\alpha$-stable errors (see [178]). Although those models are nicely LAN, with root-$n$ contiguity rates, classical estimators fail miserably. Much efforts have been made, in the literature on extremes, to renormalize them in some appropriate fashion; but the result is not even rate-optimal! Starting from a LAD preliminary, [178] proposes $R$-estimators based on stable scores that remain root-$n$ consistent under the whole family of stable distributions, irrespective of their asymmetry and tail index. With the LAD estimator, those $R$-estimators are the only rate-optimal ones available in the literature.

### 1.5 Depth- and quantile-based inference

In a pathbreaking *Annals of Statistics* paper with discussion ([158], [159]) Marc (with Davy Paindaveine and Miroslav Šiman) establishes an unexpected link between the concept of halfspace depth introduced by Tukey, and a directional version of Koenker and Bassett’s celebrated concept of quantile regression. That connection brings to halfspace depth the benefits, inherited
from the $L_1$ nature of quantiles, of such results as a Bahadur representation and the asymptotic normality of depth hyperplane coefficients, as well as those of linear programming computation. The same approach also can be adopted in multiple-output regression, where it produces nested depth regression tubes (see [189]) with an interpretation of multiple-output quantile regression surfaces. The concept is particularly attractive in the construction of multivariate growth charts (see [170] for applications) and multivariate outlier detection. Indeed, medical doctors typically consult single-output growth charts which only can diagnose marginal outliers, while a nowhere marginally outlying observation clearly can be a multivariate outlier.

A more direct approach to a multiple-output generalization of Koenker and Bassett’s concept is taken in [198], where ellipsoids, rather than hyperplanes, are characterized via the minimization of the expected check function, leading to elliptical linear regression quantiles. The delicate issue there consists in showing that the resulting minimization can be turned into a convex optimization problem.

In a somewhat different direction, [188] is introducing a new type of quantile-related periodogram where the traditional least squares regression of the observations $X_t$ on sines and cosines is replaced with quantile regression in the Koenker and Bassett spirit, yielding a distinct cross-periodogram for every couple $(\tau_1, \tau_2)$ of quantile orders. This leads to a rank-based periodogram kernel with interesting properties and suggests the definition of a population copula-based cross-spectral density kernel. Contrary to traditional spectral densities, copula-based spectra do not require any moment conditions to exist, and are able to account for any features of the bivariate joint distributions of the couples $(X_t, X_{t-k})$, $k \in \mathbb{Z}$. Rank-based periodograms and copula-based cross-spectral densities are invariant under continuous order-preserving marginal transformations of the series under study, so that the approach very neatly separates the marginal and serial features of the underlying process.

Consistency and asymptotic normality results are obtained in [188] for a smoothed version of those rank-based periodograms, but only pointwise with respect to $\tau \in [0, 1]$. A closely related rank-based periodogram kernel, consistently estimating, after due smoothing, the same copula-based cross-spectral densities, is studied in [195], where uniform asymptotics are carefully established. The main difficulty there lies in the fact that the ranks involved are not computed from exchangeable observations. In [204] and [206], locally stationary versions of the same spectral concepts, requiring the definition of a new concept of local stationarity, are introduced.

1.6 A measure-transportation approach to multivariate distribution and quantile functions, ranks and signs

More recently, Marc ([199], with V. Chernozhukov, A. Galichon, and M. Henry and [216] with E. del Barrio, J. Cuesta-Albertos, and C. Matrán) proposed a new concept of statistical depth, based on Monge-Kantorovich measure transportation ideas. In [216] and [217], this approach is developed, from a statistical decision theory point of view, into a general theory of center-outward distribution and quantile functions, the empirical versions of which yield multivariate concepts of ranks and signs.

This new concept relies on a monotone (in the sense of gradients of convex functions) probability integral transformation to the uniform distribution over the unit ball. That transformation is entirely canonical, and, contrary to halfspace depth contours (which are always convex), its contours account for the “shape” of the underlying distribution—yielding banana- and pear-shaped contours for banana- and pear-shaped distributions, respectively. In the particular case of univariate or spherical/elliptical distributions, that concept produces ranks and signs that co-
incide with the usual and well-accepted ones, paving the way to a general theory of rank-based inference in multivariate analysis.

Those ranks and signs, unlike the many other concepts (componentwise ranks, spatial ranks, Mahalanobis ranks, ...) that can be found in the literature, enjoy all the fundamental structural properties\(^1\) that make traditional ranks a fundamental concept in statistical decision theory and a successful inferential tool. Empirical and population center-outward distribution functions moreover are related by a Glivenko-Cantelli property—the quintessential property of their traditional univariate counterparts. Similarly, the center-outward quantile functions, in sharp contrast with all other concepts developed so far, produce quantile contours and regions with preassigned probability contents that do not depend on the actual underlying distribution.


2 Nonparametric analysis of high-dimensional time-series data

Another important domain of Marc’s research activity in nonparametric statistics is the analysis of high-dimensional time series. Such time series appear in a huge variety of applications, but Marc’s contributions were motivated, mainly, by his collaboration with econometricians, both in macroeconometrics (M. Forni, M. Lippi, L. Reichlin, P. Zaffaroni) and in finance (M. Barigozzi, D. Veredas, S. Soccorsi, C. Trucios). Datasets in those areas typically come under the form of very large panels of interrelated time series; “large” here means several hundreds to one thousand. A parametric approach in dimension \(n\), even for the simple VAR(1) model, requires \(n(3n + 1)/2\) parameters, that is, for \(n = 1,000\), a hopeless parameter space of dimension 1,500,500 ... A nonparametric approach is thus the only reasonable attitude.

Inspired by Brillinger’s concept of dynamic principal components, [109] (more than 1,900 Google Scholar citations) introduces a generalized factor model (GDFM) method by which the spectral density matrix of the \(n\)-dimensional process under study is decomposed into two components: the “common component”, of low rank and with exploding (as \(n \to \infty\)) dynamic eigenvalues\(^2\), and the “idiosyncratic component”, with uniformly bounded dynamic eigenvalues. The intuitive interpretation is that the “common component” is driven by a small number of shocks (the “market shocks”) that are hitting almost all time series in the panel and cause

---

\(^1\)All those properties originate in the fact that the sigma-field of the observations factorizes into the product of the sub-sigma-field of (residual) ranks and signs and the sub-sigma-field of the (residual) order statistic. The latter is sufficient and complete for the nuisance (the unspecified density of the noise driving the model); the former is essentially maximal ancillary (as shown in [216]); by Basu’s theorem, they are mutually independent.

\(^2\)Those dynamic eigenvalues are those of spectral density matrices; they are functions of the frequency.
exploding eigenvalues, whereas the “idiosyncratic” shocks only affect a small number of series, yielding mild cross-correlations hence bounded dynamic eigenvalues.

That approach is essentially model-free (on this particular point, see [183]); apart from second-order stationarity, the existence of spectral densities, and the presence of a finite (but unspecified) number of diverging spectral eigenvalues, no constraints are put on the data-generating process, and the GDFM decomposition follows as a canonical representation result. This is in sharp contrast with other factor model decompositions considered in the literature, where the covariance matrix $\Gamma_0$ is assumed to decompose into the sum of a low-rank matrix plus a sparse one—an assumption which is static (it does not take into account the serial dependence features of the data), and does not even resist affine transformations.

That GDFM method (the first paper [109] describing the approach attracted more than 2,000 Google Scholar citations) had a significant impact in the econometric community.

In a series of papers joint with Marco Forni, Mario Lippi, and Lucrezia Reichlin, a consistent estimation procedure is constructed ([109]), consistency rates are studied ([128]), and a forecasting strategy is proposed ([136]); [114], [117] and [127] establish the applicability of the method on real macroeconomic data. The problem of determining the number of independent common shocks underlying the data is treated by Marc and Roman Liška [146], based on information criterion techniques; some financial applications are considered in [163], [167], and [229].

Up to that point, the methods, based on Brillinger’s theory, which involves two-sided filters, had relatively poor performances at the end of the observation period, hence was of little help in forecasting problems. That problem found an elegant solution in [193] and [205] which, building on recent results for reduced-rank stochastic processes, proposes and studies a strictly one-sided alternative to the original Forni-Hallin-Lippi-Reichlin (2000) method.

In his ongoing research, Marc (in collaboration with Matteo Barigozzi) is developing general dynamic factor model methods for a nonparametric study of volatilities in high dimension ([196], [201], [202], [208], [212], [220]). The locally stationary case is considered in [215]. Extensions to time series of functional data have been obtained ([223]) with, as a first step, a functional extension of Brillinger’s theory of dynamic principal components ([192], in collaboration with Siegfried Hörmann).

Dynamic factor models with large cross-sectional dimension are attracting increasing attention in finance and macroeconomic applications. In finance, they are at the heart of the extensions proposed by Chamberlain, Rothschild, and Ingersol of the classical arbitrage pricing theory. In macroeconomics, they are used to identify economy-wide and global shocks, to construct coincident indexes, and to forecast individual macroeconomic quantities by taking advantage of the information scattered in a huge number of related series.

The Forni-Hallin-Lippi-Reichlin dynamic factor methods are currently implemented by a number of economic and financial institutions, including several central banks and national statistical offices, who are using it in their current analysis of the business cycle (among them, the European Central Bank, the Federal Reserve, the National Bank of Switzerland, the Banca d’Italia, ...). A real-time coincident indicator of the EURO area business cycle (EuroCOIN), based on the same method, is published monthly by the London-based Center for Economic Policy Research and the Banca d’Italia: see http://www.cepr.org/data/EuroCOIN/. Also based on that method, a similar monthly index is established for the US economy by the Federal Reserve of Chicago.

1972


1973


1976


1977


1978


1980


1981


1982


1983


1984


1985


1986


1987


1988


1989


1990


1991


1992


1993


1994


1995


1996


1997


1998


1999


2000


2001


2002


2003


2004


2005


2006


2007


2008


2009


2010


2011


2012


2013


2014


2015


2016


2017


2018


2019


2020


Submitted — Preprints


Special Issues and Books Edited


