Statistical procedures based on empirical characteristic functions

Zdeněk Hlávka and Marie Hušková

Charles University, Prague

ROBUST 2016

Outline

Introduction

Well-known from basic courses:

There is a one-to-one relationship between distribution function and characteristics function

X – random variable

$F(x) = P(X \leq x)$, $x \in \mathcal{R}$ – distribution function

$\varphi(t) = E(\exp\{itX\})$, $t \in \mathcal{R}$ – characteristic function

Statistical problems typically formulated in terms of distribution functions and their parameters, therefore also in terms of characteristics functions.

$$\varphi(t) = E(\exp\{itX\}) = C(t) + iS(t)$$

Goodness-of-fit tests

• **Goodness-of-fit tests**, simplest formulation:
$X_1, \ldots, X_n$ are i.i.d. random variables with d.f. F

$H_0 : F = F_0$ for a given F_0

against

$H_1 : H_0$ is not true

More often:

$H_0^* : F \in \mathcal{F}$, $\mathcal{F}$ a system of distributions, typically depending on parameters—nuisance parameters

Kolmogorov-Smirnov type tests

Typical test procedures for H_0 versus H_1 are based on empirical distribution functions

$$\widehat{F}_n(x) = \frac{1}{n}\sum_{j=1}^{n} I\{X_i \le x\}, \quad x \in \mathbb{R}$$

Kolmogorov-Smirnov test: $\sup_{x\in\mathbb{R}} |\widehat{F}_n(x) - F_0(x)|$

Cramér-von-Mises test: $\int_{x\in\mathbb{R}} |\widehat{F}_n(x) - F_0(x)|^2 dF_0(x)$

Anderson-Darling test: $\int_{x\in\mathbb{R}} |\widehat{F}_n(x) - F_0(x)|^2 w(x) dF_0(x)$

$w(x)$- weight function, often $w(x) = (F_0(x)(1 - F_0(x)))^{-1}$

χ^2- test

Advantage: if F_0 is continuous, the distribution of KS and CVM under H_0 does not depend on F_0 (distribution free test statistics).

Similar problem:

(i) H_0^S: distribution F is symmetric ($F(x) = 1 - F(x)\,\forall x$),

(ii) two sample tests—two independent samples, we are testing that they have the same distribution,

(iii) independence tests,

(iv) change-point tests.

Empirical characteristic function based procedures

$X_1, \ldots, X_n$ — i.i.d. random variables

Testing problem H_0 versus H_1 can be equivalently expressed as

$H_0 : \varphi = \varphi_0$ for a given φ_0 versus $H_1 : H_0$ is not true

$\varphi(u) = E \exp\{iuX_j\}$, $u \in \mathcal{R}$ – characteristic function (CF)

$\widehat{\varphi}_n(u) = \frac{1}{n} \sum_{j=1}^{n} \exp\{iuX_j\}$, $u \in \mathcal{R}$ – empirical characteristic function (ECF)

Test statistic:

$$T_n(w) = \int_{\mathcal{R}} |\widehat{\varphi}_n(u) - \varphi_0(u)|^2 w(u) du$$

$w(\cdot)$- weight function (usually, nonnegative, symmetric).

- Large values indicate that the null hypothesis is violated.
- Question is critical value – simulation (F_0 given), asymptotics for $n \to \infty$, simulated critical values, bootstrap.
- Noticing that $\exp\{iuX_j\} = \cos(uX_j) + i\sin(uX_j)$, $u \in \mathcal{R}$ and by symmetry of $w(\cdot)$ we get

$$T_n(w) = \int_{\mathcal{R}} \Big(\frac{1}{n} \sum_{j=1}^{n} \big(U_j(u) - E_0 U_j(u) \big) \Big)^2 \times w(u) du = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{v=1}^{n} J_w(X_j - X_v)$$

$J_w(x) = \int_{\mathcal{R}} \cos(ux) w(u) du$ and $E_0(...)$ denotes the expectation under the null hypothesis and

$$U_j(u) = \cos(uX_j) + \sin(uX_j), \quad u \in \mathcal{R}.$$

Asymptotic behavior of $T_n(w)$:
Under the null hypothesis and $\int_{\mathcal{R}} u^2 w(u)du < \infty$

$$nT_n(w) \to^d \int_{\mathcal{R}} V^2(u)w(u)du,$$

where $\{V(u);\ u \in \mathcal{R}\}$ is a Gaussian process with zero mean and

$$cov(V(u_1), V(u_2)) = cov_0(U_j(u_1), U_j(u_2))$$

$cov_0(.)$ – covariance under the null hypothesis. Generally,

$$nT_n(w) \to^d \int_{\mathcal{R}} \left(\tilde{V}(u) - \sqrt{n}(EU_j(u) - E_0U_j(u))\right)^2 \quad w(u)du.$$

If $\int \left(EU_j(u) - E_0U_j(u)\right)^2 dw(t) > 0$ then

$$nT_n(w) \to^P \infty.$$

Something from the history

H. Cramér (1946) – classical book, empirical characteristic function mentioned

Feuerverger and Mureika (1997), Annals of Statistics

Sandor Csörgő (1984) – Proceedings of Asymptotic Statistics, 1984, Praha

Ushakov (1999) – Selected Topics in Characteristics Functions (book)

Meintanis (2016), South African Statistical Journals – survey paper with discussions

More general setup:

Klebanov (2005) – N-distances and Their Applications (book)

Procedures based on Probability generating function – see talk of Hudecová

Rizzo and Székely et al (2010,...)

Further procedures based on empirical characteristic functions for various statistical problems in recent years:

- tests for symmetry,
- test for independence,
- two-sample problem,
- change point problem,
- nuisance parameters,

asymptotics, computational aspects, simulations, applications.

Two-sample problem

$Y_1, \ldots, Y_n$ – independent random variables
F_j – distribution function of Y_j
Testing problem

$$H_0 : F_1 = \ldots = F_n$$
$$H_1 : F_1 = \ldots = F_m \neq F_{m+1} = \ldots = F_n \qquad \text{for} \quad m < n,$$

F_1 and F_n are unknown, m - known.

$$T_{m,n-m}(w) = \frac{m(n-m)}{n} \int_{-\infty}^{\infty} |\widehat{\varphi}_m(t) - \widehat{\varphi}^0_{n-m}(t)|^2 w(t) dt,$$

$w(\cdot)$ is a nonnegative weight function,
$\widehat{\varphi}_m(t)$ and $\widehat{\varphi}^0_{n-m}(t)$ – empirical characteristic functions based on $Y_1, \ldots, Y_m$ and $Y_{m+1}, \ldots, Y_n$, respectively, i.e.,

$$\widehat{\varphi}_m(t) = \frac{1}{k} \sum_{j=1}^{m} \exp\{itY_j\}, \quad \widehat{\varphi}^0_{n-m}(t) = \frac{1}{n-m} \sum_{j=m+1}^{n} \exp\{itY_j\}.$$

Under the null hypothesis

$$ET_{m,n-m}(w) = \frac{m(n-m)}{n} \int_{-\infty}^{\infty} E|\widehat{\varphi}_m(t) - \widehat{\varphi}^0_{n-m}(t)|^2 w(t)dt$$

$$= \frac{m(n-m)}{n} \int_{-\infty}^{\infty} E\Big(\frac{1}{m}\sum_{j=1}^{m} U_j(t) - \frac{1}{n-m}\sum_{j=m+1}^{n} U_j(t)\Big)^2 w(t)dt$$

$$= \int_{-\infty}^{\infty} var(U_1(t))w(t)dt.$$

Generally,

$$E\Big(\frac{1}{m}\sum_{j=1}^{m} U_j(t) - \frac{1}{n-m}\sum_{j=m+1}^{n} U_j(t)\Big)^2$$

$$= \frac{var(U_1(t)}{m} + \frac{var(U_n(t)}{n-m} + \Big(EU_1(t) - EU_1(t)\Big)^2.$$

Limit behavior under H_0 and $0 < \int_{\mathcal{R}} t^2 w(t)dt < \infty$:
For $m = m_n$, $m_n/n \to \theta_0 \in (0,1)$

$$T_{m,n-m}(w) \to^d \int_{\mathcal{R}} V^2(t)w(t)dt,$$

$\{V(t);, t \in \mathcal{R}\}$ – Gaussian process with zero mean and covariance structure

$$cov(V(t_1), V(t_2)) = cov(U_j(t_1), U_j(t_2)).$$

For testing — null hypothesis rejected for large values of test statistic, approximation for critical values — either simulation of the limit distribution with estimated covariance, or some bootstrap.

Consistent test.

Multivariate version — quite straightforward.

Simulations

NOTATION USED IN THE TABLES

N: $N(0,1,0)$
N1: $N(0.4,1,0)$
N2: $N(0.7,1,0)$
N3: $N(0,1.5,0)$
N4: $N(0,2,0)$
N5: $N(0,1,0.6)$
N6: $N(0,1,0.9)$
MN1: $MN(0.2,1,0)$
MN2: $MN(0.4,1,0)$
MN3: $MN(0,1.2,0)$
MN4: $MN(0,1.5,0)$
MN5: $MN(0,1.2,0.5)$
MN6: $MN(0,1.2,0.8)$
Γ1: $\Gamma(0.01,1)$
Γ2: $\Gamma(0.5,0.5)$
Γ3: $\Gamma(0.5,1.0)$
Γ4: $\Gamma(1.0,1.0)$.

F_2	T_1	$T_{1.5}$	T_2	$\tau_{1.5}$	τ_2	τ_4	T_1	$T_{1.5}$	T_2	$\tau_{1.5}$	τ_2	τ_4
N	5 10	5 10	6 10	6 11	6 11	6 10	5 10	5 10	5 10	5 10	5 10	6 10
N1	24 35	26 36	27 37	22 33	25 35	28 39	41 53	43 55	45 56	37 48	40 51	46 55
N2	68 78	71 80	72 81	62 73	68 78	75 82	92 96	93 96	94 97	88 93	91 95	94 97
N3	30 43	30 44	30 44	27 41	29 43	29 44	52 67	54 68	55 68	48 63	52 66	54 69
N4	61 74	62 75	63 76	56 69	61 73	62 77	90 96	92 96	92 97	87 94	90 96	93 97
N5	7 12	6 12	6 12	7 12	7 12	6 11	7 13	6 12	6 12	7 14	6 13	6 10
N6	8 15	7 14	7 13	10 18	8 15	6 11	10 18	6 12	6 12	15 25	11 18	6 12
MN1	11 17	11 18	11 18	9 16	10 17	10 18	14 23	15 23	15 23	14 22	14 23	16 24
MN2	27 37	28 39	29 40	23 33	26 37	31 40	43 54	45 56	46 57	39 51	43 54	48 59
MN3	9 16	9 16	9 17	9 15	8 15	9 15	14 22	14 22	14 22	14 21	14 22	14 23
MN4	18 27	18 27	18 27	17 24	18 26	17 26	30 39	31 40	31 40	29 38	30 40	30 41
MN5	9 17	9 17	9 17	9 15	8 16	9 15	14 22	14 23	14 22	14 21	14 23	14 22
MN6	10 18	10 17	10 17	9 16	9 16	9 15	15 23	15 23	15 23	15 23	15 23	14 23

Table 1 ($d = 2$): Percentage of rejection of the null hypothesis with F_1 the standard multivariate normal distribution based on samples of size $n_1 = n_2 = 25$ (left part) and $n_1 = n_2 = 50$ (right part). Nominal size: $\alpha = 5\%$ (left entry), $\alpha = 10\%$ (right entry)

Change-point problem

$Y_1, \ldots, Y_n$ – independent random variables
F_j – distribution function of Y_j

$$H_0 : F_1 = \ldots = F_n$$
$$H_1 : F_1 = \ldots = F_m \neq F_{m+1} = \ldots = F_n \qquad \text{for} \quad m < n,$$

m, F_1 and F_n are unknown.

$$T_{n,\gamma}(w) = \max_{1 \leq k < n} \left(\frac{k(n-k)}{n^2} \right)^{\gamma} \frac{k(n-k)}{n} \int_{-\infty}^{\infty} |\widehat{\varphi}_k(t) - \widehat{\varphi}^0_{n-k}(t)|^2 w(t) dt, \tag{1}$$

$w(\cdot)$ is a nonnegative weight function, $\gamma \in (0, 1)$
$\widehat{\varphi}_k(t)$ and $\widehat{\varphi}^0_{n-k}(t)$ – empirical characteristic functions based on $Y_1, \ldots, Y_k$ and $Y_{k+1}, \ldots, Y_n$, respectively.

Under H_0

$$E\Big[\Big(\frac{k(n-k)}{n^2}\Big)^{\gamma}\frac{k(n-k)}{n}\int_{-\infty}^{\infty}|\widehat{\varphi}_k(t)-\widehat{\varphi}^0_{n-k}(t)|^2 w(t)dt\Big]$$

$$=\Big(\frac{k(n-k)}{n^2}\Big)^{\gamma}\int_{-\infty}^{\infty} var(Z_j(t))w(t)dt.$$

Generally,

$$E\Big[\Big(\frac{k(n-k)}{n^2}\Big)^{\gamma}\frac{k(n-k)}{n}\int_{-\infty}^{\infty}|\widehat{\varphi}_k(t)-\widehat{\varphi}^0_{n-k}(t)|^2 w(t)dt\Big]$$

$$=\Big(\frac{k(n-k)}{n^2}\Big)^{\gamma}\int_{-\infty}^{\infty}\Big(\dots$$

Limit behavior under H_0 and $0 < \int_{\mathcal{R}} t^2 w(t)dt < \infty$:

For $m = m_n$, $m_n/n \to \theta_0 \in (0,1)$

$$T_{n,\gamma}(w) \to^d \sup_{s\in(0,1)} (s(1-s))^{\gamma-1} \int_{\mathcal{R}} Z^2(s,t)w(t)dt,$$

$\gamma \in (0,1]$, $\{V(s,t);\ s \in (0,1),\ t \in \mathcal{R}\}$ – Gaussian process with zero mean and covariance structure $(0 < s_1 \le s_2 < 1)$

$$cov(Z(s_1,t_1), Z(s_2 t_2)) = s_1(1-s_2)cov(U_j(t_1), U_j(t_2)).$$

Some theoretical results

We are interested in limit behavior ($n \to \infty$) of

$$\sup_{s\in(0,1)} (s(1-s))^{\gamma-1} \int_{\mathcal{R}} (Z_n(s,u) - sZ_n(1,u))^2 w(u)du$$

$\gamma \in (0,1]$

$$Z_n(s,\ u) = \frac{1}{\sqrt{n}} \sum_{k=1}^{\lfloor sn \rfloor} (U_j(u) - EU_j(u)), \quad u \in \mathcal{R}, \quad s \in (0,1)$$

$$U_j(u) = \cos(Y_j u) + \sin(Y_j u).$$

The following holds true:

a) For any $0 < s < 1$ it holds $\sup_n E \int_{\mathcal{R}} (Z_n(s,u))^2 w(u) du < \infty$.

b) There exists an $a > 0$, $0 < D < \infty$ such that for any $0 < s < 1$ it holds

$$\sup_n E \left| Z_n^2(s,u_1) - Z_n^2(s,u_2) \right| \leq D \|u_1 - u_2\|^a.$$

c) The marginal distributions of $\{Z_n(s,u)\}$ converge to the marginal distributions of a Gaussian process $\{Z(s,u)\}$ with covariance structure $(0 < s_1 \leq s_2 < 1)$

$$cov\{Z(s_1,u_1), Z(s_2,u_2)\} = s\, cov(U_1(u_1), U_1(u_2)).$$

Then

$$\int_{\mathbb{R}} (Z_n(s,u))^2 w(u) du \to^d \int_{\mathbb{R}} (Z(s,u) - sZ(1,u))^2 w(u) du$$

for any fixed $s \in (0,1)$ by Theorem 22 in Ibragimov and Chasminkij (1981)

Still needed to investigate

$$X_n(s) = \left(\int_{\mathcal{R}} (Z_n(s, \boldsymbol{u}) - sZ_n(1, u))^2 w(u) du \right)^{1/2}, \quad s \in (0, 1)$$

it means to prove tightness and convergence of the finite dimensional distribution.

Procedures when nuisance parameters are present

Linear models

$Y_1, \ldots, Y_n$ – independent observations following the linear model

$$Y_j = \mathbf{x}_j^{\mathrm{T}} \beta + c e_j, \ j = 1, 2, ..., n,$$

$\mathbf{x}_j = (1, x_{j2}, ..., x_{jp})^{\mathrm{T}} \in R^p, \ j = 1, 2, ..., n$ – known regressors,

$\beta \in R^p$ and $c > 0$ – unspecified regression and scale parameters,

$e_j, \ j = 1, 2, ..., n$ – errors assumed to be i.i.d. random variables having distribution function $F(\cdot)$.
We wish to test the null hypothesis

$$H_0 : F \equiv F_0,$$

against general alternatives.

$\widehat{e}_j = (Y_j - \mathbf{x}_j^{\mathrm{T}} \widehat{\boldsymbol{\beta}}_n)/\widehat{c}_n, \ j = 1, 2, ..., n$ – residuals

Nonparametric version

Model: (X, Y) are observed

$$Y = m(X) + \sigma(X)e, \tag{2}$$

$m(\cdot)$ and $\sigma(\cdot)$ – unspecified regression and scale functions,

e – error with a distribution function F, characteristic function $\varphi(t)$, mean zero and unit variance,

To test the null hypothesis

$$H_0 : F \in \mathcal{F} = \{F_\vartheta,\ \vartheta \in \Theta\}$$

$\mathcal{F}$ – parametric family of distributions indexed by $\vartheta \in \Theta \subseteq \mathbb{R}^q,\ q \geq 1$.

$$H_0 : \varphi \in \{\varphi(\cdot;\vartheta), \vartheta \in \Theta\},$$

$\varphi(\cdot;\vartheta)$ – characteristic function corresponding to F_ϑ, for some (unspecified) $\vartheta \in \Theta$.

The proposed test statistic based on the residuals

$$\widehat{e}_j = (Y_j - \widehat{m}_n(X_j))/\widehat{\sigma}_n(X_j),\ j = 1, 2, ..., n, \tag{3}$$

$\widehat{m}_n(\cdot)$ and $\widehat{\sigma}_n^2(\cdot)$ – kernel estimators of $m(\cdot)$ and $\sigma^2(\cdot)$, the corresponding empirical characteristic function (ECF):

$$\varphi_n(t) = \frac{1}{n}\sum_{j=1}^{n} e^{it\hat{e}_j}.$$

Test statistic:

$$T_{n,w} = n\int_{-\infty}^{\infty} |\varphi_n(t) - \varphi(t;\widehat{\vartheta}_n)|^2 w(t)dt \tag{4}$$

$\widehat{\vartheta}_n$ – a suitable estimator of ϑ,

$w(\cdot)$ – a symmetric nonnegative weight function.

$(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. random vectors such that

$$Y_j = m(X_j) + \sigma(X_j)e_j, \quad j = 1, \ldots, n, \tag{5}$$

where $e_1, \ldots, e_n$, $X_1, \ldots, X_n$, $m(.)$ and $\sigma(.)$ satisfy:

- (A.1) Let $e_1, \ldots, e_n$ be i.i.d. random variables with zero mean, unit variance and $Ee_j^4 < \infty$ and characteristic function $\varphi(t; \vartheta)$, $t \in \mathbb{R}^1$, where $\vartheta = (\vartheta_1, \ldots, \vartheta_q)^T \in \Theta$, ϑ_0 denotes the true parameter value.
- (A.2) On the real and imaginary parts of $\varphi(t; \vartheta)$ denoted by $C(t; \vartheta)$ and $S(t; \vartheta)$, we assume that the first partial derivatives w.r.t. t as well as $\vartheta_1, \ldots, \vartheta_q$ exist. Particularly, we assume that $\dot{C}_s(t, \vartheta) = \frac{\partial C(t,\vartheta)}{\partial \vartheta_s}$, $\dot{S}_s(t, \vartheta) = \frac{\partial S(t,\vartheta)}{\partial \vartheta_s}$, $s = 1, \ldots, q$, are bounded continuous in ϑ in a neighborhood of ϑ_0 (which is the true parameter value), for each t. The first derivatives $C'(t; \vartheta)$ and $S'(t; \vartheta)$ w.r.t. t are bounded and continuous for all t in a neighborhood of ϑ_0.

- (A.3) $X_1, \ldots, X_n$ are i.i.d. on $[0,1]$ with common positive continuous density f_X.
- (A.4) Let $(e_1, \ldots, e_n)$ and $(X_1, \ldots, X_n)$ be independent.
- (A.5) Let $m(\cdot)$ and $\sigma(\cdot)$ be functions on $[0,1]$ with Lipschitz first derivative, $\sigma(x) > 0$, $x \in [0,1]$
- (A.6) The weight function w is nonnegative and symmetric, and

$$\int_{-\infty}^{\infty} t^4 w(t) dt < \infty.$$

- (A.7) Let $\widehat{\vartheta}_n$ be an estimator of ϑ_0 such that

$$\sqrt{n}(\widehat{\vartheta}_n - \vartheta_0) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \psi(e_j; \vartheta_0) + o_P(1)$$

$\psi(z;\vartheta) = (\psi_1(z;\vartheta), \ldots, \psi_q(z;\vartheta))^T$ are continuously differentiable functions w.r.t. to z and continuous in components of ϑ in a neighborhood of ϑ_0 and such that $E_\vartheta \psi(e_j;\vartheta) = \mathbf{0}$ and $E_\vartheta ||\psi(e_j;\vartheta)||^2 < \infty$ for ϑ in a neighborhood of ϑ_0

Estimators of $m(.), \sigma(.)$ are kernel type generated by the kernel $K(\cdot)$ and the bandwidth $h = h_n$ satisfying

- (A.8) Let K be a symmetric twice continuously differentiable density on $[-1, 1]$ with $K(-1) = K(1) = 0$.
- (A.10) Let $\{h_n\}$ be a sequence of the bandwidth such that $\lim_{n\to\infty} nh_n^2 = \infty$ and $\lim_{n\to\infty} nh_n^{3+\delta} = 0$ for some $\delta > 0$.

We use the following estimators of the density function $f_X(.)$ of X_j's, regression function $m(.)$ and variance function $\sigma^2(.)$:

$$\widehat{f}_X(x) = \frac{1}{nh_n} \sum_{j=1}^{n} K((X_j - x)/h_n), \quad \widehat{m}_n(x) = \frac{1}{nh_n \widehat{f}_X(x)} \sum_{j=1}^{n} K((X_j - x)/h_n) Y_j,$$

$$\widehat{\sigma}_n^2(x) = \frac{1}{nh_n \widehat{f}_X(x)} \sum_{j=1}^{n} K((X_j - x)/h_n)(Y_j - \widehat{m}_n(x))^2, \quad x \in [0, 1].$$

Recall that the residuals $\widehat{e}_j$ are defined above. Choice of the estimators $\widehat{\vartheta}_n$ of ϑ_0 satisfying (A.8) – for maximum likelihood type estimator $\widetilde{\vartheta}_n$

$$\sqrt{n}(\widetilde{\vartheta}_n - \vartheta_0) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \mathbf{h}(e_j; \vartheta_0) + o_P(1)$$

for a measurable $\mathbf{h}(.; \vartheta_0)$.
e_j's are replaced by the respective residuals $\widehat{e}_j$ we get for the respective estimator $\widehat{\vartheta}_n$ (ass.A.8) holds true with

$$\boldsymbol{\psi}(x; \vartheta) = \mathbf{h}(x; \vartheta) - xE_\vartheta \mathbf{h}'(e_1; \vartheta) + \frac{x^2 - 1}{2} E_\vartheta e_1 \mathbf{h}'(e_1; \vartheta), \quad x \in \mathbb{R}^1.$$

Theorem *Let assumptions (A.1)-(A.10) be satisfied. Then under the null hypothesis, as* $n \to \infty$,

$$T_{n,w} \to^d \int ||Z_0(t)||^2 w(t) dt,$$

where $\{Z_0(t),\ t \in \mathbb{R}^1\}$ *is a Gaussian process with the covariance structure of the form* $cov(Z_0(t_1), Z_0(t_2))$ where $Z_0(t) := Z(t; \vartheta_0)$ with

$$\begin{aligned} Z(t;\vartheta) = \quad & \big(\cos(te_1) + \sin(te_1) - E_\vartheta(\cos(te_1) + \sin(te_1))\big) \\ & -te_1(C(t,\vartheta) - S(t,\vartheta)) - t\frac{e_1^2 - 1}{2}(C'(t,\vartheta) + S'(t,\vartheta)) \\ & -\psi^T(e_1;\vartheta)(\dot{\mathbf{C}}(t;\vartheta) + \dot{\mathbf{S}}(t;\vartheta)). \end{aligned}$$

• The explicit form of the limit distribution of $T_{n,w}$ is unknown even under the null hypothesis. It depends on the hypothetical distribution of the error terms and the chosen estimator of the nuisance parameter ϑ.
• Surprisingly it does not depend on the density f_X of X_i's, the functions $m(.)$ and $\sigma(.)$ and even not on the kernel $K(.)$ and the bandwidth h_n.
• The limit distribution does not provide an approximation for the critical values. However, a special parametric bootstrap does it.
• The crucial part of proof is on the process

$$Z_n(t;\widehat{\vartheta}_n) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \left\{ \sin(t\hat{e}_j) + \cos(t\hat{e}_j) - C(t;\widehat{\vartheta}_n) - S(t;\widehat{\vartheta}_n) \right\}, \quad t \in \mathbb{R}^1,$$

behaves asymptotically as the Gaussian process $\{Z_0(t); t \in \mathbb{R}^1\}$ described above.

Bootstrap

The parametric bootstrap Neumeyer et al. (2006) .
!) bootstrap errors $e^*_{n1}, \ldots, e^*_{nn}$ – a random sample of size n from the distribution $F(.; \widehat{\vartheta}_n)$,
2) The bootstrap observations :

$$Y^*_{nj} = \widehat{m}(X_j) + e^*_{nj}\widehat{\sigma}_n(X_j),\ j = 1, \ldots, n,$$

3) The bootstrap version $T^*_{n,w}$ of the test statistic is defined as $T_{n,w}$ with $Y_1, \ldots, Y_n$ replaced by $Y^*_{n1}, \ldots, Y^*_{nn}$ and $\widehat{\vartheta}_n$ is replaced by its bootstrap counterpart.

It can be shown that

(i) under H_0 and ass. (A.1) – (A.10), given $Y_1, \ldots, Y_n$ the limit distribution of $T^*_{n,w}$ is the same limit distribution as that of $T_{n,w}$,
(ii) under alternatives plus some assumptions $T^*_{n,w} = O_{P^*}(1)$ holds true in probability ($P^*(.)$ denotes conditional probability given $Y_1, \ldots, Y_n$.

Simulations

Model:

$$Y_j = \beta_0 + \beta_1 X_j + \beta_2 X_j^2 + e_j, \quad j = 1, \ldots, n,$$

X_j – i.i.d. uniform on (0,1),

$\beta_0 = 0 \quad \beta_1 = \beta_2 = 1,$

$w(t) = \exp\{-\gamma t^2\}, \quad \forall t \quad \gamma > 0,$

5000 replications, bootstrap size $B = 100$,

distribution of the errors:

normal(N), Laplace (LP), $\beta(1, \vartheta)$, χ^2_θ, t_ϑ, skewnormal(SN_ϑ), asymmetric Laplace (AL), logistic (LG).

	$n = 25$						$n = 50$					
	KS	CM	$\gamma = 0.1$	0.5	0.75	1.0	KS	CM	$\gamma = 0.1$	0.5	0.75	1.0
N	5 11	5 10	5 10	5 10	5 10	5 10	5 10	5 10	5 10	5 10	6 11	5 11
LP	19 28	22 32	20 29	26 35	27 37	27 37	34 47	43 55	41 53	46 57	45 56	44 55
LG	8 14	9 15	8 14	11 18	12 19	13 20	10 17	12 20	11 18	16 24	17 25	17 25
$S^0_{1.5}$	41 49	46 53	42 51	51 59	53 60	53 60	66 73	73 78	70 76	77 82	78 83	78 83
$S^0_{1.75}$	20 27	23 29	20 27	27 34	28 35	29 36	32 40	37 45	34 41	44 51	46 53	47 53
$S^{-1}_{1.5}$	48 57	54 62	49 58	62 69	63 71	64 71	76 83	84 88	80 85	89 92	90 93	90 93
$S^{-1}_{1.75}$	22 30	26 33	22 30	31 39	33 40	34 41	40 49	45 53	40 48	52 60	55 63	56 64
$b_{0.5}$	40 54	50 64	53 66	50 66	44 61	38 56	77 87	89 95	92 96	92 97	89 95	84 94
$b_{0.75}$	16 28	23 36	29 41	21 36	14 28	10 21	38 53	55 70	64 77	60 77	48 70	36 60
χ^2_3	41 54	51 64	45 58	61 73	63 75	63 75	73 84	86 92	83 90	92 96	93 96	93 97
χ^2_5	27 39	33 46	28 40	42 54	44 57	45 58	51 64	64 76	58 71	76 85	79 87	79 87
χ^2_7	21 31	25 36	21 32	31 43	33 45	34 46	38 52	49 62	43 56	62 72	64 75	65 76
t_3	24 33	29 37	25 34	34 43	36 44	37 44	42 53	52 60	48 57	58 66	59 67	58 67
t_4	15 23	18 26	16 23	23 31	24 33	25 33	26 36	33 42	30 39	40 50	41 51	42 52
t_5	11 19	14 22	12 19	18 25	19 27	20 28	17 26	22 31	19 28	28 37	30 39	30 40

Percentage of rejection for the normality null hypothesis at level 5% (left entry), 10% (right entry)

	$n = 25$						$n = 50$					
	KS	CM	$\gamma = 0.1$	0.5	0.75	1.0	KS	CM	$\gamma = 0.1$	0.5	0.75	1.0
LP	5 11	5 11	5 10	5 10	5 10	5 10	5 11	6 11	5 10	5 10	5 10	5 10
$S^0_{1.5}$	18 25	21 27	9 16	17 24	20 27	22 30	25 33	29 37	15 22	25 33	29 37	32 40
$S^0_{1.75}$	8 14	9 15	7 13	9 15	8 15	9 15	10 17	11 18	11 19	12 20	11 19	12 18
$S^{-1}_{1.5}$	32 44	38 49	19 29	35 48	38 50	40 51	61 74	68 80	42 56	70 81	72 83	73 83
$S^{-1}_{1.75}$	12 19	13 21	10 18	14 24	14 23	14 22	22 34	25 37	17 28	27 42	27 41	27 40
$b_{0.5}$	30 48	35 56	55 69	64 80	47 71	29 52	71 86	81 93	93 97	98 99	96 99	88 98
$b_{0.75}$	10 20	12 28	36 51	44 64	23 48	7 23	26 47	40 67	76 87	91 97	82 95	55 87
χ^2_3	31 47	35 52	29 42	44 61	42 60	37 56	66 82	75 88	66 79	88 95	86 94	82 93
χ^2_5	18 32	20 35	19 31	28 46	25 42	22 37	42 62	48 69	43 60	68 84	65 82	58 78
χ^2_7	14 25	14 26	16 27	21 36	17 32	15 27	30 49	34 54	33 49	55 72	49 70	42 65
SN_3	8 16	7 16	11 20	13 25	10 21	7 16	16 30	17 33	23 38	35 54	28 48	21 39
SN_6	14 26	14 28	19 30	25 41	19 35	15 29	34 52	38 58	41 57	62 79	55 75	46 68
SN_{10}	16 30	18 33	22 34	30 47	23 42	19 35	41 60	46 66	50 66	72 86	66 82	56 76
$AL_{0.4}$	39 56	45 62	33 46	50 67	50 67	47 64	78 89	85 93	74 85	91 96	91 97	89 96
$AL_{0.6}$	26 38	28 42	20 30	30 44	29 44	28 43	56 71	61 76	47 61	65 78	64 77	62 76

Percentage of rejection for the Laplace null hypothesis at level 5% (left entry), 10% (right entry)

$\gamma =$	$n = 50$				$n = 100$			
	0.1	0.5	0.75	1.0	0.1	0.5	0.75	1.0
$AL_{0.4}$	5 10	5 10	5 10	4 10	4 10	5 10	4 9	4 9
$AL_{0.5}$	5 10	5 10	6 12	6 12	5 10	6 11	5 11	4 10
$AL_{0.75}$	5 10	5 9	4 9	3 7	6 11	5 10	5 10	4 9
AL_2	5 10	5 10	6 12	6 13	5 9	5 11	6 11	5 11
AL_3	5 9	5 10	5 9	4 10	5 10	5 10	5 9	4 9
AL_4	5 10	4 9	4 8	3 8	5 10	5 9	4 9	4 9
χ^2_1	15 21	65 75	83 89	85 91	23 31	93 96	98 99	98 99
χ^2_2	6 11	9 15	13 25	17 30	6 11	15 26	24 39	25 42
LN_1	8 14	22 33	47 60	56 70	9 14	51 64	81 89	86 93
$LN_{1.5}$	11 17	27 40	56 72	75 85	12 18	73 83	96 98	99 100
$T_{0.75}$	7 13	14 22	24 37	30 44	8 13	21 34	46 62	55 71
T_1	9 15	23 35	48 62	56 71	9 14	50 64	81 89	86 92
$W_{0.5}$	16 22	60 71	83 91	90 95	25 32	95 97	99 100	100 100
$W_{0.75}$	10 16	47 58	70 79	75 84	13 19	82 89	95 97	96 98

Percentage of rejection for the asymmetric Laplace null hypothesis at level 5% (left entry), 10% (right entry)

Computations

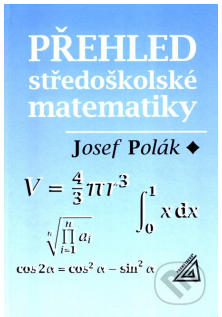
Computations for ECF-based statistics

AIM:

Advantages vs. disadvantages

CONCLUSION: The ECF test statistic has computationally expensive closed form expression.

Bibliography



Polák (2005). *Přehled středoškolské matematiky*, Prométheus.

Henze, Hlávka & Meintanis (2014). Testing for spherical symmetry via the empirical characteristic function. *Statistics* 48(6), 1282–1296.

Meintanis & Hlávka (2010). Goodness-of-fit tests for bivariate and multivariate skew-normal distributions. *Scandinavian Journal of Statistics* 37(4), 701–714.

Computations for ECF-based statistics

Typically, research papers say that ECF based test statistics, e.g,

$$T = \int_{-\infty}^{\infty} |\hat{\varphi}_X(t) - \hat{\varphi}_Y(t)|^2 w(t) dt,$$

can be (it is easy to see, using simple algebra, clearly) expressed as

$$T = \frac{1}{n^2} \sum_{i,j} I_w(X_i - X_j) + \frac{1}{m^2} \sum_{i,j} I_w(Y_i - Y_j) - \frac{2}{mn} \sum_{i,j} I_w(X_i - Y_j),$$

where, for example, $I_w(D) = \sqrt{\pi} \exp(-D^2)$ or $2/(1 + D^2)$.

Computations for two-sample problem

In the two-sample problem, we use the test statistic

$$T = \int_{-\infty}^{\infty} |\hat{\varphi}_X(t) - \hat{\varphi}_Y(t)|^2 w(t) dt,$$

where $\hat{\varphi}_X(t) = \frac{1}{n} \sum \exp(itX_i)$ and $\hat{\varphi}_Y(t) = \frac{1}{m} \sum \exp(itY_i)$.

Let us recall some helpful formulas:

$$\begin{aligned} |x + iy| &= \sqrt{x^2 + y^2}, \\ \exp(it) &= \cos(t) + i \sin(t). \end{aligned}$$

It follows that T is equal to

$$\int_{-\infty}^{\infty} |\frac{1}{n} \sum \{\cos(tX_i) + i \sin(tX_i)\} - \frac{1}{m} \sum \{\cos(tY_i) + i \sin(tY_i)\}|^2 w(t) dt.$$

Next, using the formula for absolute value, we have

$$\begin{aligned} T &= \int_{-\infty}^{\infty} \left[\left\{ \frac{1}{n} \sum \cos(tX_i) - \frac{1}{m} \sum \cos(tY_i) \right\}^2 \right. \\ &\quad \left. + \left\{ \frac{1}{n} \sum \sin(tX_i) - \frac{1}{m} \sum \sin(tY_i) \right\}^2 \right] w(t)dt \\ &= \int_{-\infty}^{\infty} \left[-\frac{2}{mn} \sum\sum \{\cos(tX_i)\cos(tY_j) + \sin(tX_i)\sin(tY_j)\} \right. \\ &\quad + \frac{1}{n^2} \sum\sum \{\cos(tX_i)\cos(tX_j) + \sin(tX_i)\sin(tX_j)\} \\ &\quad \left. + \frac{1}{m^2} \sum\sum \{\cos(tY_i)\cos(tY_j) + \sin(tY_i)\sin(tY_j)\} \right] w(t)dt \end{aligned}$$

Using

$$\cos(\alpha - \beta) = \cos(\alpha)\cos(\beta) + \sin(\alpha)\sin(\beta),$$

we obtain

$$
\begin{aligned}
T &= \int_{-\infty}^{\infty} \Big[-\frac{2}{mn} \sum_{i,j} \cos\{t(X_i - Y_j)\} + \frac{1}{n^2} \sum_{i,j} \cos\{t(X_i - X_j)\} \\
&\quad + \frac{1}{m^2} \sum_{i,j} \cos\{t(Y_i - Y_j)\} \Big] w(t)dt \\
&= -\frac{2}{mn} \sum_{i,j} \int \cos\{t(X_i - Y_j)\} w(t)dt + \frac{1}{n^2} \int \sum_{i,j} \cos\{t(X_i - X_j)\} w(t)dt \\
&\quad + \frac{1}{m^2} \sum_{i,j} \int \cos\{t(Y_i - Y_j)\} w(t)dt
\end{aligned}
$$

and it remains to choose the weight function $w(t)$ so that $\int \cos\{tD\} w(t)dt$ has closed form expression.

Favorite choices are $w(t) = \exp(-at^2)$ or $w(t) = \exp(-b|t|)$ because

$$\begin{aligned} \int_{-\infty}^{\infty} \cos\{tD\} \exp(-at^2) dt &= \sqrt{\frac{\pi}{a}} \exp(-D^2/4a), \\ \int_{-\infty}^{\infty} \cos\{tD\} \exp(-b|t|) dt &= \frac{2b}{b^2 + D^2}. \end{aligned}$$

The resulting algorithm is:

1. Calculate the $(n+m)^2$ differences D_{ij}^{XX}, D_{kl}^{YY}, and D_{ik}^{XY}.
2. Calculate the integrals $I_{ij} = \int \cos\{tD_{ij}\} w(t) dt$.
3. Calculate T as the (weighted) sum of the integrals I_{ij}.

```
T1=0
for (i in 1:n) {
    for (j in 1:n) {
        T1=T1+iw(x[i]-x[j])
    }
}
T2=0
for (i in 1:m) {
    for (j in 1:m) {
        T2=T2+iw(y[i]-y[j])
    }
}
T3=0
for (i in 1:n) {
    for (j in 1:m) {
        T3=T3+iw(x[i]-y[j])
    }
}
T=T1/(n^2)+T2/(m^2)-2*T3/(n*m)
```

Speed of calculation

Good news: ECF lead to closed form expression.

Bad news: the algorithm is not fast (we have to calculate and sum $n(n-1)/2 + m(m-1)/2 + n * m = (n+m)(n+m-1)/2$ terms).

Naive R implementation of the two-sample ECF test statistics leads to:

$n = m = 100$ 0.04s

$n = m = 200$ 0.18s

$n = m = 400$ 0.68s

$n = m = 800$ 2.74s

Other testing problems

This method can be used in other testing problems, for example:

k-sample problem $H_0 : \varphi_1 = \cdots = \varphi_k$,

goodness-of-fit $H_0 : \varphi_X = \varphi$ or some other property of φ,

multivariate symmetry $H_0 : \varphi(t) = \Phi(\|t\|^2)$,

independence $H_0 : \varphi(t, s) = \varphi(t)\varphi(s)$.

In the following, we shortly discuss a *change-point problem* (generalization of the two-sample problem):

H_0 : Y_i are iid

vs.

H_1 : $\exists k$ such that $Y_1, \ldots, Y_k \sim F_1$ and $Y_{k+1}, \ldots, Y_T \sim F_2$.

Two-sample changepoint (and bootstrap)

We need to compare samples $Y_1, \ldots, Y_k$ and $Y_{k+1}, \ldots, Y_T$ for all $k = 1, \ldots, n-1$.

The ECF test statistic is

$$T = \max_k \gamma(k) \int_{-\infty}^{\infty} |\hat{\varphi}_k(t) - \hat{\varphi}^{k+1}(t)|^2 w(t) dt,$$

where $\hat{\varphi}_k(t) = \frac{1}{k} \sum_{i=1}^{k} \exp(itY_i)$, $\hat{\varphi}^{k+1}(t) = \frac{1}{T-k} \sum_{i=k+1}^{T} \exp(itY_i)$ and $\gamma(k)$ is a weight function.

Critical values are typically obtained by bootstrap leading to computational difficulties.

Speed of calculation

Naive R implementation of the two-sample test statistics leads to:

$n = m = 100$ 0.04s

$n = m = 200$ 0.18s

$n = m = 400$ 0.68s

$n = m = 800$ 2.74s

In changepoint analysis with bootstrap critical values (say $B = 1000$), we need to calculate this roughly $BT = B(n + m)$ times leading to:

$n = m = 100$ 8000s = 2.2h

$n = m = 200$ 72000s = 20h

$n = m = 400$ 544000s = 6.3d

$n = m = 800$ 4384000s = 51d

Speed of calculation

Higher speed is possible by using C code and compiled shared library:

```
twosam <- function (x,y) {.C("twosam",x=as.double(x),
    y=as.double(y),n=as.integer(length(x)),
    m=as.integer(length(y)),t=double(1))$t
}
dyn.load("./twosam.so")
T2=twosam(x,y)
```

The computation time for $n = m = 800$ is reduced from 2.74s (corresponding to 51 days) to 0.008s (corresponding to 21 minutes).

The code can be further optimized by using some simple relations.

```
void twosam(double *x, double *y, int *n, int *m, double *t)
{   int i, j;
    double t1, t2, t3;
    t1 = 0.0; t2 = 0.0; t3 = 0.0;
    for(i = 0; i < *n; i++)
    {
        for( j = 0; j < *n; j++)
        {
            t1 += 1.0 / (1.0 + ((x[j]-x[i]) * (x[j]-x[i])));
        }
    }
    for(i = 0; i < *m; i++)
    {
        for( j = 0; j < *m; j++)
        {
            t2 += 1.0 / (1.0 + ((y[j]-y[i]) * (y[j]-y[i])));
        }
    }
    for(i = 0; i < *n; i++)
    {
        for( j = 0; j < *m; j++)
        {
            t3 += 1.0 / (1.0 + ((y[j]-x[i]) * (y[j]-x[i])));
        }
    }
    *t = t1/(*n * *n) +  t2/(*m * *m) - 2.0 * t3/(*n * *m) ;
}
```

Advantages

Advantages:

1. closed form expression,
2. easy generalization to more dimensions.

Recall that multivariate CF is $\varphi(t) = E \exp\{it^\top X\}$.

All derivations for multivariate ECFs are very similar.

Multivariate setup

In the two-dimensional two-sample problem, we use the same test statistic

$$T = \int_{-\infty}^{\infty} |\hat{\varphi}_X(t) - \hat{\varphi}_Y(t)|^2 w(t) dt,$$

with multivariate ECFs $\hat{\varphi}_X(t) = \frac{1}{n} \sum \exp(it^\top X_i)$ and $\hat{\varphi}_Y(t) = \frac{1}{m} \sum \exp(it^\top Y_i)$ leading to

$$\begin{aligned} T \;=\; & -\frac{2}{mn} \sum_{i,j} \int \cos\{t^\top (X_i - Y_j)\} w(t) dt + \frac{1}{n^2} \int \sum_{i,j} \cos\{t^\top (X_i - X_j)\} w(t) dt \\ & + \frac{1}{m^2} \sum_{i,j} \int \cos\{t^\top (Y_i - Y_j)\} w(t) dt \end{aligned}$$

Multivariate setup

Using $\cos(\alpha+\beta) = \cos(\alpha)\cos(\beta) - \sin(\alpha)\sin(\beta)$, we have

$$\begin{aligned}
&\int \cos\{t^\top (X_i - Y_j)\} w(t) dt \\
&= \int \cos\{t_1(X_{i1} - Y_{j1}) + t_2(X_{i2} - Y_{j2})\} w(t) dt \\
&= \int [\cos\{t_1(X_{i1} - Y_{j1})\} \cos\{t_2(X_{i2} - Y_{j2})\} \\
&\quad - \sin\{t_1(X_{i1} - Y_{j1})\} \sin\{t_2(X_{i2} - Y_{j2})\}] w_1(t_1) w_2(t_2) dt_1 dt_2 \\
&= \int \cos\{t_1(X_{i1} - Y_{j1})\} w_1(t_1) dt_1 \int \cos\{t_2(X_{i2} - Y_{j2})\} w_2(t_2) dt_2
\end{aligned}$$

if $w(t) = w_1(t_1) w_2(t_2)$, where $w_i(x)$ are symmetric.

Multivariate setup

The resulting expression for the two-dimensional test statistics

$$\begin{aligned} T &= -\frac{2}{mn}\sum_{i,j} I_w(X_{i1}-Y_{j1})I_w(X_{i2}-Y_{j2}) \\ &\quad + \frac{1}{n^2}\sum_{i,j} I_w(X_{i1}-X_{j1})I_w(X_{i2}-X_{j2}) \\ &\quad + \frac{1}{m^2}\sum_{i,j} I_w(Y_{i1}-Y_{j1})I_w(Y_{i2}-Y_{j2}). \end{aligned}$$

is not much more complicated than in one-dimension because it only replaces the terms $I_w(X_{i1}-Y_{j1})$ by $I_w(X_{i1}-Y_{j1})I_w(X_{i2}-Y_{j2})$.

Disadvantages

Disadvantages:

1. choice of tuning parameters (of the weight function),
2. nuisance parameters (bootstrap),
3. computationally intensive (but some tests of this type are even worse).

ECF-based test of spherical symmetry

Let $\varphi(t) = E(\exp(it^\top X)),\ t \in \mathbb{R}^p$, denote the characteristic function (CF) of random vector X.

$\mathcal{H}_0$: there is some function $\phi : \mathbb{R} \to \mathbb{R}$ such that $\varphi(t) = \phi(\|t\|^2),\ t \in \mathbb{R}^p$.

Test statistic can be based on discrepancies

$$D_n(t, s) = \hat{\varphi}_n(t) - \hat{\varphi}_n(s),$$

computed on pairs of points $t, s \in \mathbb{R}^p$ such that $\|t\| = \|s\|$.

ECF-based test of spherical symmetry

A Cramér-von Mises type test statistic is

$$\mathrm{CM}_n = n \int_0^\infty \left(\sum_{j=1}^{K} \sum_{m=1}^{K} |D_n(\rho u_j, \rho u_m)|^2 \right) W(\rho) d\rho,$$

where u_i, $i = 1, \ldots, K$, are points scattered on unit sphere.

Straightfoward algebra yields:

$$\mathrm{CM}_n = \frac{1}{n} \sum_{r,s=1}^{K} \sum_{l,m=1}^{n} \left[I_W(u_r^\top X_{lm}) + I_W(u_s^\top X_{lm}) - 2 I_W(u_s^\top X_l - u_r^\top X_m) \right],$$

where $X_{lm} = X_l - X_m$ and $I_W(z) := \int_0^\infty \cos(\rho z)\, W(\rho) d\rho$.

MGF-based test of skew-normality

Moment generating function of bivariate skew-normal distribution satisfies:

$$\delta_2 \frac{\partial M(t_1,t_2)}{\partial t_1} - \delta_1 \frac{\partial M(t_1,t_2)}{\partial t_2} = \left[(\delta_2 - \omega\delta_1)t_1 - (\delta_1 - \omega\delta_2)t_2\right] M(t_1,t_2)$$

Test statistics:

$$T_{n,W}(\hat{\vartheta}_n) = n \int_{\mathbf{R}^2} D_n^2(t_1,t_2;\hat{\vartheta}_n) W(t_1,t_2) dt_1 dt_2,$$

where $D_n(t_1,t_2;\vartheta)$ is

$$\delta_2 \frac{\partial M_n(t_1,t_2)}{\partial t_1} - \delta_1 \frac{\partial M_n(t_1,t_2)}{\partial t_2} - \left[(\delta_2 - \omega\delta_1)t_1 - (\delta_1 - \omega\delta_2)t_2\right] M_n(t_1,t_2).$$

MGF-based test of skew-normality

By straightforward algebra (it is easy to see, clearly)

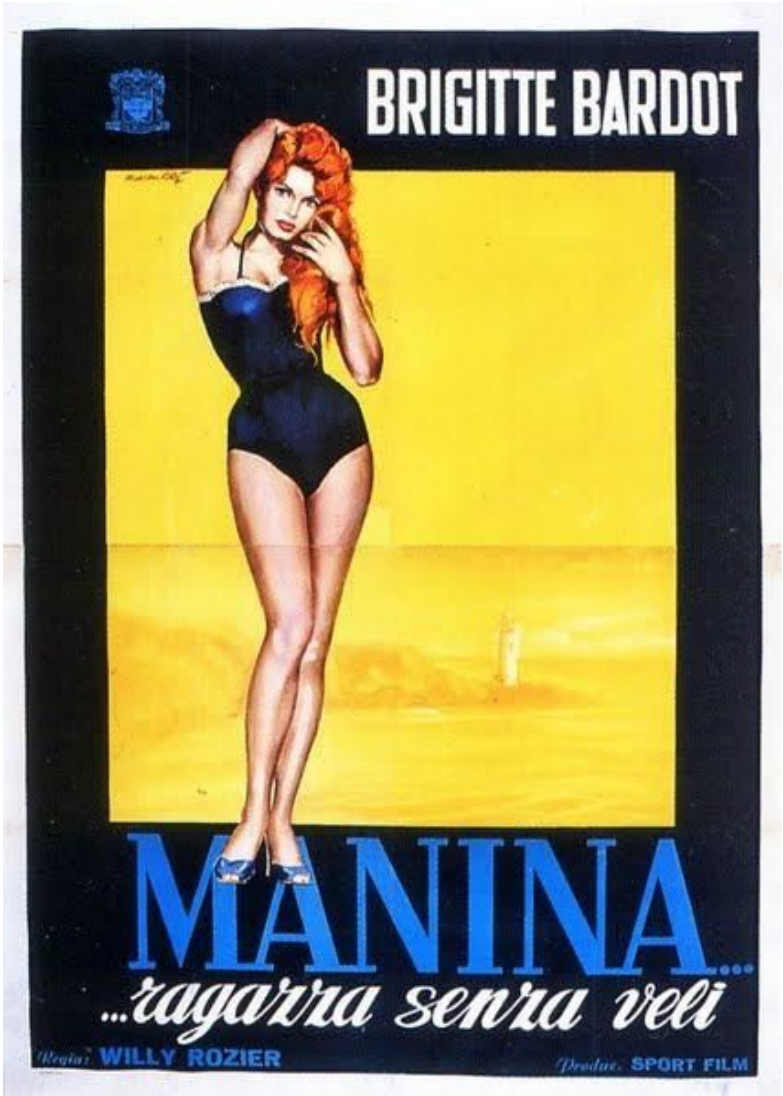
$$T_{n,w}(\vartheta) = \frac{1}{n}\sum_{j,k=1}^{n}\left[\delta_2^2 X_{1j}X_{1k} + \delta_1^2 X_{2j}X_{2k} - 2\delta_1\delta_2 X_{1j}X_{2k}\right] I_0(X_{1jk})I_0(X_{2jk})$$

$$+\frac{1}{n}\sum_{j,k=1}^{n}\left[\kappa_2^2 I_2(X_{1jk})I_0(X_{2jk}) + \kappa_1^2 I_2(X_{2jk})I_0(X_{1jk}) - 2\kappa_1\kappa_2 I_1(X_{1jk})I_1(X_{2jk})\right]$$

$$+\frac{2}{n}\sum_{j,k=1}^{n}\{[\delta_2\kappa_1 X_{1j} - \delta_1\kappa_1 X_{2j}]I_1(X_{2jk})I_0(X_{1jk})$$

$$+[\delta_1\kappa_2 X_{2j} - \delta_2\kappa_2 X_{1j}]I_1(X_{1jk})I_0(X_{2jk})\}$$

where $X_{mjk} = X_{mj} + X_{mk}$, $m = 1, 2$, and $I_m(z) = \int_{-\infty}^{\infty} t^m e^{tz} w(t)dt$.


BRIGITTE BARDOT
MANINA...
...ragazza senza veli
Regia: WILLY ROZIER
Produz. SPORT FILM

Tests for martingale difference hypothesis (MDH)

Testing procedures which detect if the observed time series is **martingale difference sequence (MDH)**

Tests detection of change-points in the conditional expectation of the series given its past.

New test statistics based on Fourier-type conditional expectations.

The asymptotic properties, simulations, applications to the real data.

Motivation for our test is from Bierens (1982).

Formulation

The standard formulation of the MDH:

$$E(Y_t|\mathbb{I}_{t-1}) = 0, \qquad t = 1, \ldots, \tag{6}$$

$\mathbb{I}_t$ – the information set available at time t, and

Y_t – represents first differences of a process which under this hypothesis forms a martingale sequence.

Standard assumption statistical models used in finance and economics: The efficient market hypothesis states that in efficient markets, prices follow a martingale and always fully and instantaneously reflect all available relevant information consisting of past prices and returns, asset returns in an efficient market.

The basic idea for the MDH is the unpredictability of macro and financial series on the basis of currently available information.

Testing for zero autocorrelation – 1978 – Ljung and Box (1978)

Bierens – 1982,

Hong 1999, Escanciano and Velasco (2006), Jong (1996)

Lobato – 2002

Escanciano and Lobato (2009)- survey

MDH for exchange rates, for instance, Belaire-Franch and Contreras (2011), Yilmaz (2003), Hong and Lee (2003), Fong et al. (1997), and Fong and Ouliaris (1995).

Less standard areas for MDH:

electricity prices (Veka, 2013)

CO2 emissions (Daskalakis et al., 2009, Charles et al., 2011a)

Null hypothesis and test statistics

$$H_0^{(1)}: \quad E(Y_t|\mathbb{I}_{t-1}) = 0, \qquad t = 1, \ldots, \tag{7}$$

$\mathbb{I}_t$ – the information set available at time t against

$$H_1^{(1)}: \quad E(Y_t|\mathbb{I}_{t-1}) = g(Y_{t-1}, \ldots, Y_{t-m}),$$
$$P(g(Y_{t-1}, \ldots, Y_{t-m}) = 0) < 1,$$

g – an arbitrary unknown function g, $m > 0$ – a chosen time-lag.

Change point version – k_0 – unknown change point

$$H_0^{(2)}: \quad E(Y_t|\mathbb{I}_{t-1}) = 0,$$
$$H_1^{(2)}: \quad E(Y_t|\mathbb{I}_{t-1}) = 0, \quad t < k_0,$$
$$\text{but } E(Y_t|\mathbb{I}_{t-1}) = g(Y_{t-1}, \ldots, Y_{t-m}), \quad t \geq k_0$$
$$P(g(Y_{t-1}, \ldots, Y_{t-m}) = 0) < 1.$$

Test procedures based on characterization (Bierens (1982)):

$$E(Y|\boldsymbol{X}) = 0 \Leftrightarrow E(Y \exp\{i\boldsymbol{X}'\boldsymbol{u}\}) = 0 \quad \boldsymbol{u} \in R^m.$$

Define:

$$S_t^{(m)}(\boldsymbol{u}) = \frac{1}{\sqrt{n}} \sum_{\tau=m+1}^{t} Y_\tau e^{i\mathbf{u}'\boldsymbol{Y}_{\tau,m}}, \quad t = m+1, \ldots, n, \tag{8}$$

$$S_t^{(m)}(u) = 0, \quad t = 0, 1, \ldots, m,$$

$\boldsymbol{Y}_{t,m} = (Y_{t-1}, Y_{t-2}, ..., Y_{t-m})'$,

$m > 0$ denotes a chosen time-lag.

Consider the integrated process

$$Q_m(s) = \int_{\mathbb{R}^m} |\frac{1}{\sqrt{n}} \sum_{\tau=\lfloor sn \rfloor+1}^{n} Y_\tau e^{i\mathbf{u}' \mathbf{Y}_{\tau,m}}|^2 w(\mathbf{u}) d\mathbf{u}, \quad 0 \le s \le 1, \qquad (9)$$

$w(\cdot)$ – a weight function.

The null hypothesis $H_0^{(1)}$ against alternative $H_1^{(1)}$ rejected if

$$T_n^{(1)} := Q_m(0) \qquad (10)$$

is large.

The null hypothesis $H_0^{(2)}$ is rejected in favor of alternative $H_1^{(2)}$ if

$$T_n^{(2)}(\gamma) := \max_{m+1 \le k \le n} Q_m(k/n)/q(k/n, \gamma) \qquad (11)$$

is large, where

$$q(s, \gamma) = (1-s)^\gamma, s \in (0,1), \quad 0 \le \gamma < 1. \qquad (12)$$

Behavior under the null hypothesis

Theorem $\{Y_t\}$ is a martingale difference sequence as well as stationary, ergodic with $E|Y_1|^{2+\delta} < \infty$ for some $\delta > 0$
$w(\cdot)$ be a measurable non-negative function on $\mathbb{R}^m$

$$w(\boldsymbol{t}) = w(-\boldsymbol{t}) > 0, \quad \textit{for} \quad \textit{all} \quad \boldsymbol{t} \in \mathbb{R}^m, \quad 0 < \int_{\mathbb{R}^m} w(\boldsymbol{t}) d\boldsymbol{t} < \infty.$$

Then as $n \to \infty$:

$$(a) \quad T_n^{(1)} \to^d \int_{\mathbb{R}^m} |Z(0, \mathbf{u})|^2 w(\mathbf{u})\, du,$$

$$(b) \quad T_n^{(2)}(\gamma) \to^d \sup_{0<s<1} \frac{1}{(1-s)^\gamma} \int_{\mathbb{R}^m} |Z(s, \mathbf{u}) - Z(1, \mathbf{u})|^2 w(\mathbf{u})\, d\mathbf{u},$$

$0 \le \gamma < 1$, $\{Z(s, \mathbf{u}), s \in [0,1], \mathbf{u} \in \mathbb{R}^m\}$ is a Gaussian process with expectation zero and covariance $(0 \le s_1 \le s_2 \le 1)$

$$cov\{Z(s_1, \boldsymbol{u}_1), Z(s_2, \boldsymbol{u}_2)\} = s_1 E\Big(Y_{m+1}^2 h(\boldsymbol{Y}_{m+1}, \boldsymbol{u}_1) h(\boldsymbol{Y}_{m+1}, \boldsymbol{u}_2)\Big), \quad \boldsymbol{u}_1, \boldsymbol{u}_2,$$

$$h(\boldsymbol{Y}_m, \boldsymbol{u}) = cos\left(\sum_{q=1}^{m} u_q Y_{m+1-q}\right) + sin\left(\sum_{q=1}^{m} u_q Y_{m+1-q}\right), \qquad (13)$$

Here $\boldsymbol{u} = (u_1, \ldots, u_m)'$, $\boldsymbol{Y}_{m+1} = (Y_m, \ldots, Y_1)'$.

The assertion of our theorem remains true if $cov\{Z(s_1, \boldsymbol{u}_1), Z(s_2, \boldsymbol{u}_2)\}$ are replaced by their consistent estimators.

Critical values can be obtained by simulating the limit distribution. But more convenient is a proper bootstrap.

Alternatives

$H_0^{(1)}$ **versus** $H_1^{(1)}$

$$Y_k = \xi_k + g(\boldsymbol{\xi}_k),$$

$\{\xi_t\}$ is a stationary and ergodic martingale difference sequence and g is a measurable function such that for some $\delta > 0$

$$P(g(\boldsymbol{\xi}_{m+1}) = 0) < 1, \qquad E|\xi_1|^{2+\delta} < \infty, \quad E|g(\boldsymbol{\xi}_{m+1})|^2 < \infty.$$

Change-point alternative with an MDS before the change $H_0^{(2)}$ **and** $T_n^{(2)}$**:**

$$Y_k = \xi_k + g(\boldsymbol{\xi}_k)\, 1_{\{k>k_0\}}, \quad k_0 = \lfloor \lambda n \rfloor$$

for some $0 < \lambda < 1$, where $(\{\xi_t\}, g)$ fulfill above.

Both tests are consistent, even sensitive w.r.t. local alternatives.

Estimator of the change point k_0:

$$\widehat{k}(\gamma) = \min\{m < k < n;\ \tilde{Q}_m(k/n)/\tilde{q}(k/n,\gamma) = \max_{m<j<n} \tilde{Q}_m(j/n)/\tilde{q}(j/n,\gamma)\},$$

$$\tilde{Q}_m(s) = \int_{\mathbb{R}^m} |S^{(m)}_{\lfloor sn \rfloor}(\boldsymbol{u}) - sS^{(m)}_n(\boldsymbol{u})|^2 w(\boldsymbol{u}) d\boldsymbol{u}.$$

Wild bootstrap

(B.1): $\{\eta_i\}_i$ are i.i.d. with mean zero, unit variance and $E\,|\eta_1|^{2+\delta} < \infty$ for some $\delta > 0$,

(B.2): $\{\eta_i\}_i$ and $\{Y_i\}_i$ are independent sequences of random variables.

Bootstrap statistics:

$$S_t^{(m)*}(\mathbf{u}) = \frac{1}{\sqrt{n}} \sum_{\tau=m+1}^{t} Y_\tau \exp\left(i\mathbf{u}'\boldsymbol{Y}_{\tau,m}\right)\eta_\tau,$$

define $T_n^{(j)*}$ analogously to $T_n^{(j)}$ with $S_t^{(m)}(\mathbf{u})$ replaced by $S_t^{(m)*}(\mathbf{u})$.

Under the null hypothesis and local alternatives:

$$P(T_n^{(j)*} \le x | Y_1, \ldots, Y_n) - P(T_n^{(j)} \le x) \to^p 0, \quad j = 1, 2, \quad x \in \mathbb{R}^1.$$

Under fixed alternatives for all x:

$$|P(T_n^{(1)*} \le x | Y_1, \ldots, Y_n) - P(\int_{\mathbb{R}^m} |Z^0(0, \mathbf{u})|^2 w(\mathbf{u})\, du \le x)| \to^p 0,$$

$\{Z^0(s, \mathbf{u}), s \in [0,1], \mathbf{u} \in \mathbb{R}^m\}$ is a Gaussian process with expectation zero and covariance $(0 \le s_1 \le s_2 \le 1)$

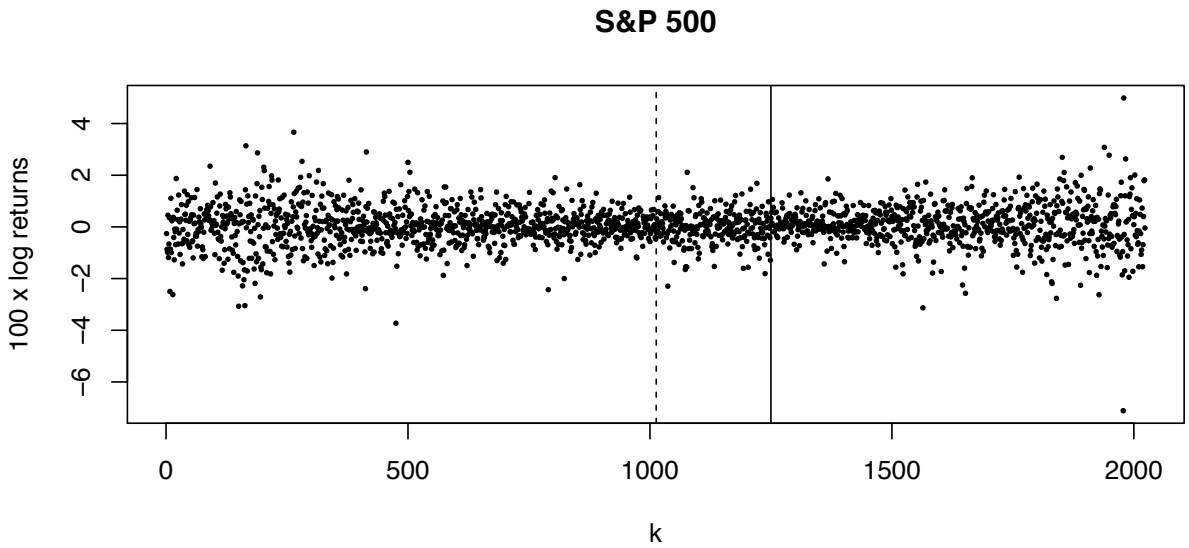
$$cov\{Z^0(s_1, \boldsymbol{u}_1), Z(s_2, \boldsymbol{u}_2)\} = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1+m}^{\lfloor ns_1 \rfloor} E\Big(Y_j^2 h(\boldsymbol{Y}_j, \boldsymbol{u}_1) h(\boldsymbol{Y}_j, \boldsymbol{u}_2)\Big), \quad \boldsymbol{u}_1, \boldsymbol{u}_2.$$

Martingale difference hypothesis (MDH)

Most efficiency studies on financial markets focus on a weak form of market efficiency through the MDH, whereby the profit expected from an asset (which is forecasted to have its future price equal to its the current price) is equal to zero.

Apart of testing **MDH** in a given time period ($H_0^{(1)}$), we test also the hypothesis of **no change** in the martingale difference structure ($H_0^{(2)}$).

Real data example: Daily scaled log returns of S&P 500 from 1990 until 1997 (source: Yahoo! Finance) have been previously analyzed by EV2006 [Escanciano and Velasco: Generalized spectral tests for the martingale difference hypothesis. *J.Econometr. 134 (2006) 151–185*].



Daily scaled log returns of S&P 500. Dashed line denotes January 1st, 1994, solid line denotes December 8th, 1994.

EV2006 conclude that MDH is not rejected for the first period (Jan1990–Dec1993) and it is rejected for the second period (Jan1994–Dec1997).

Change-point analysis

We obtain the change–point estimate $\widehat{k} = 1250$ corresponding to a change occurring on December 8th, 1994.

We obtain p-value 0.649 for data observed until December 7th, 1994, and p-value 0.000 for data observed from December 8th, 1994, which implies that the MDH is not rejected for the first period (Jan1990–Dec7, 1994), while it is rejected for the second perriod (Dec8, 1994–Dec1997).

To confirm that there is no further change in the first period we tested the change-point hypothesis $H_0^{(2)}$ and obtained a p-value of 0.526.

Conclusions

The hypothesis of no change in the martingale difference structure between January 1990 and December 1997 is rejected. The change in the martingale difference structure of the S&P 500 log returns occurred in December 1994, almost one year later than the change-point considered previously in EV2006.

MDH is not rejected for log returns until December 7th, 1994, and it is rejected for log returns observed after December 8th, 1994.

The hypothesis of no change in the martingale difference structure is not rejected using the data between January 1990 and December 7th, 1994.

Economic crises in 1990s

- Japanese asset price bubble (1986–2003)
- Bank stock crisis (Israel 1983)
- Black Monday (1987)
- Savings and loan crisis of the 1980s and 1990s in the U.S.
- Early 1990s Recession
- 1991 India economic crisis
- Finnish banking crisis (1990s)
- Swedish banking crisis (1990s)
- 1994 Tequila crisis in Mexico
- 1997 Asian financial crisis
- 1998 Russian financial crisis
- Argentine economic crisis (1999–2002)

Source: wikipedia

Economic crises in 1990s

- Japanese asset price bubble (1986–2003)
- Bank stock crisis (Israel 1983)
- Black Monday (1987)
- Savings and loan crisis of the 1980s and 1990s in the U.S.
- Early 1990s Recession
- 1991 India economic crisis
- Finnish banking crisis (1990s)
- Swedish banking crisis (1990s)
- **1994 Tequila crisis in Mexico**
- 1997 Asian financial crisis
- 1998 Russian financial crisis
- Argentine economic crisis (1999–2002)

Source: wikipedia

The Tequila crisis

The Tequila crisis was a currency crisis sparked by the Mexican government's sudden devaluation of the peso against the U.S. dollar in December 1994. The Mexican economy experienced hyperinflation of around 52% and mutual funds began liquidating Mexican assets as well as emerging market assets in general. The effects spread to economies in Asia and the rest of Latin America. Source: wikipedia