Nonparametric test for detecting change in distribution with panel data

Denys Pommeret, Mohamed Boutahar, Badih Ghattas

Nonparametric test for detecting change in distribution with panel data

M. Boutahar, B. Ghattas and D. Pommeret
Institute of Mathematics of Luminy.
Luminy Faculty of Sciences. 163 Av. de Luminy 13288 Marseille Cedex 9 - France
boutahar@univmed.fr , ghattas@univmed.fr , pommeret@univmed.fr

May 1, 2011

Abstract

This paper considers the problem of comparing two processes with panel data. A nonparametric test is proposed for detecting a monotone change in the link between the two process distributions. The test statistic is of CUSUM type, based on the empirical distribution functions. The asymptotic distribution of the proposed statistic is derived and its finite sample property is examined by bootstrap procedures through Monte Carlo simulations.

keywordsnonparametric estimation panel data process

1 Introduction

Many situations lead to the comparison of two random processes. In a parametric case, the problem of change detection has been widely studied in the time series literature. A common problem is to test a change in the mean or in the variance of the time series by using a parametric model (see for instance [8] or [7], and references therein). In the Gaussian case comparisons of processes are considered through their covariance structures (see [9], [12]). These distribution assumptions can be relaxed when the study concerns processes observed through panel data. This situation is frequently encountered in medical follow-up studies when two groups of patients are observed and compared. Each subject in the study gives rise to a random process (X_t) denoting the measurement of the patient up to time t (such data are referred to as panel data). In this context, [3, 2, 1] considered the problem of testing the equality of mean functions and proposed new multi-sample tests for panel count data.

In this paper we consider the general problem of comparison of two processes which may differ by a transformation of their distributions. Our purpose is to test whether this transformation changes over time. For this, two panels

are considered: $(X_{i,t})_{1\leq i\leq N_x;1\leq t\leq n}$ and $(Y_{i,t})_{1\leq i\leq N_y;1\leq t\leq n}$, not necessarily independent; that is, we can have i.i.d. paired observations $(X_i, Y_i)_{i=1,\cdots,N}$ with dependence between X_i and Y_i. It is assumed that for each t, the $X_{i,t}, 1 \leq i \leq N_x$ (resp. $Y_{i,t}, 1 \leq i \leq N_y$) are i.i.d. random variables with common distribution function F_t (resp. G_t) and with support $\mathcal{X}$ (resp. $\mathcal{Y}$). Also we assume that for all $1 \leq t \leq n$ there exists monotone transformations h_t such that the following equality in distribution holds: $X_t \stackrel{d}{=} h_t(Y_t)$. Without loss of generality we consider that the functions $h_t(.)$ are increasing. Note that if F_t is invertible then there exists a trivial transformation h_t given by $h_t = F_t^{-1} \circ G_t$. We are interested in testing whenever this transformation is time independent; that is, for all t, the equality $h_t = h$ occurs. A simple illustration is the case where X_t and Y_t are Gaussian processes with mean m_X and m_Y and variance $t\sigma_X^2$ and $t\sigma_Y^2$, respectively. In that case the function h is linear.

More generally, observing both processes X_t and Y_t with panel data we want to test

$$H_0 : \forall t, h_t = h \quad \text{against} \quad H_1 : \exists t_1 \neq t_2, h_{t_1} \neq h_{t_2}.$$

It is clear that H_0 coincides with the equality in distribution: $X_t \stackrel{d}{=} (h(Y_t))$, for all t. Following [8] (see also [7]), we construct a non parametric test statistic based on the empirical estimator of h_t, denoted by $\widehat{h}_t$. We show that $\widehat{h}_t$ is proportional to a Brownian bridge under H_0.

When H_0 is not rejected, it is of interest to estimate h and to interpret its estimator $\widehat{h}$. Then this test can be viewed as a first step permitting to legitimate estimation and interpretation of a constant transformation h between the distributions of two samples, possibly paired.

The paper is organized as follows: In Section 2 we construct the test statistic. In Section 3 we perform a simulation study using a bootstrap procedure to evaluate the finite sample property of the test. The power is evaluated against alternatives where there are smooth scale or position time changes in the process distribution. Section 4 contains brief concluding remarks.

2 The test statistic

A natural nonparametric estimator of h_t is given by

$$\widehat{h}_t(\cdot) \quad = \quad X_{(N_x \widehat{G}_t(\cdot)),t},$$

where $X_{(i),t}$ denotes the ith order statistic and $\widehat{G}_t$ is the empirical distribution function of $(Y_{i,t})_{1\leq i\leq N_y}$, that is

$$\widehat{G}_t(x) \quad = \quad \frac{1}{N_y} \sum_{i=1}^{N_y} \mathbf{1}_{\{Y_{i,t} \leq x\}}.$$

A nonparametric test is considered to test the variation of h_t. For $\tau \in (0,1)$, $x \in \mathcal{Y}$, write

$$B_n(\tau, x) = \frac{1}{\sqrt{n}\widehat{\sigma}_n}\left(\sum_{t=1}^{[n\tau]} \widehat{h}_t(x) - \frac{[n\tau]}{n}\sum_{t=1}^{n} \widehat{h}_t(x)\right), \quad (1)$$

where

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{t=1}^{n}(\widehat{h}_t(x) - \bar{h}(x))^2, \bar{h}(x) = \frac{1}{n}\sum_{t=1}^{n}\widehat{h}_t(x).$$

For a given square integrable function w we define the following test statistic

$$S_n(w) = \int_{\mathbb{R}} w(x) \sup_{1\le\tau\le 1} |B_n(\tau, x)|\, dx.$$

To establish the limiting distribution of the statistic $S_n(w)$ under the null, we need the following assumptions:

- Assumption 1. There exists $a < \infty$ such that $N_x/(N_x + N_y) \to a$.
- Assumption 2. There exist $\gamma_1 > 0$ and $\gamma_2 > 0$ such that $f_t(x) \ge \gamma_1$ and $g_t(y) \ge \gamma_2$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, where f_t and g_t are the density functions of X_t and Y_t.
- Assumption 3. For all $x \in \mathcal{X}$, there exist $0 < \overline{\sigma}_2^2(x) < \infty$ such that

$$\frac{1}{n}\sum_{t=1}^{n}\sigma_{1,t}^2(x) \to \overline{\sigma}_2^2(x), \qquad \text{as } n \to \infty,$$

  where

$$\sigma_{1,t}^2(x) = \sigma_t^2(x)\frac{N_x + N_y}{N_x N_y}, \text{ and } \sigma_t^2(x) = \frac{G_t(x)(1 - G_t(x))}{f_t^2(h_t(x))}. \quad (2)$$

- Assumption 4.

$$\frac{n(N_x + N_y)}{N_x N_y} \to 0.$$

REMARK 2.1 *Assumptions 1 and 2 are standard. Assumption 3 states that the second moments converge on average. If Assumption 1 is satisfied, Assumption 4 is equivalent to $n = o(N_x)$ or $n = o(N_y)$.*

THEOREM 2.1 *Let assumptions 1-4 hold. Then under the null H_0 we have the following convergence in distribution*

$$S_n(w) \xrightarrow{d} S(w) = B_\infty \int_{\mathbb{R}} w(x)dx, \text{ as } n \to \infty, N_x \to \infty \text{ and } N_y \to \infty, \quad (3)$$

where $B_\infty = \sup_{0\le\tau\le 1} |B(\tau)|$, and B is a Brownian bridge.

REMARK 2.2 *The cumulative distribution function of B_∞ is given by (see [4])*

$$F_{B_\infty}(z) \quad = \quad 1 + 2\sum_{k=1}^{\infty}(-1)^k \exp\{-2k^2z^2\}.$$

Before proving Theorem 1, we state three lemmas.

LEMMA 2.1 *Under Assumption 1 we have*

$$\left(\frac{N_xN_y}{N_x+N_y}\right)^{1/2}(\widehat{h}_t(x) - h_t(x)) \xrightarrow{d} N(0, \sigma_t^2(x)), \qquad \text{as } N_x \to \infty, N_y \to \infty, \quad (4)$$

where $\sigma_t^2(x)$ is given by (2).

Proof $(\mathbf{1}_{\{Y_{i,t}\leq x\}})$ is an i.i.d sequence with mean $G_t(x)$ and variance $G_t(x)(1-G_t(x))$, hence an immediate application of the central limit theorem yields

$$N_y^{1/2}\left(\widehat{G}_t(x) - G_t(x)\right) \quad \xrightarrow{d} \quad N(0, G_t(x)(1-G_t(x))). \quad (5)$$

By the delta-method the last convergence implies that

$$N_y^{1/2}\left(F^{-1}\left(\widehat{G}_t(x)\right) - F^{-1}\left(G_t(x)\right)\right) \quad \xrightarrow{d} \quad N(0, \sigma_t^2(x)). \quad (6)$$

For $p \in]0;1[$ fixed, denote by $\widehat{F}_t^{-1}(p)$ the sample p-quantile; that is, $\widehat{F}_t^{-1}(p) = X_{(r),t}$, where $r = [N_xp] + 1$. By Theorem 3 of [13] we obtain

$$N_x^{1/2}(\widehat{F}_t^{-1}(p) - F_t^{-1}(p)) \xrightarrow{d} N\left(0, \frac{p(1-p)}{f_t^2(F_t^{-1}(p))}\right), \qquad \forall p \in (0,1). \quad (7)$$

Let $\phi_X(t) = \mathbb{E}(\exp(itX))$ denotes the characteristic function of the random variable X and let $\phi_{X|Y}(t) = \mathbb{E}(\exp(itX) \mid Y)$ denotes the conditional characteristic function of the random variable X conditional on Y. We have

$$\begin{aligned}
\widetilde{H}_t &= \left(\frac{N_xN_y}{N_x+N_y}\right)^{1/2}(\widehat{h}_t(x) - h_t(x)) \\
&= \left(\frac{N_xN_y}{N_x+N_y}\right)^{1/2}\left(\widehat{F}_t^{-1}(\widehat{G}_t(x)) - F_t^{-1}(G_t(x))\right) \\
&= \widetilde{H}_{1,t} + \widetilde{H}_{2,t},
\end{aligned}$$

where

$$\widetilde{H}_{1,t} = \left(\frac{N_y}{N_x+N_y}\right)^{1/2} N_x^{1/2}\left(\widehat{F}_t^{-1}(\widehat{G}_t(x)) - F_t^{-1}(\widehat{G}_t(x))\right)$$

$$\widetilde{H}_{2,t} = \left(\frac{N_x}{N_x+N_y}\right)^{1/2} N_y^{1/2}\left(F_t^{-1}(\widehat{G}_t(x)) - F_t^{-1}(G_t(x))\right).$$

Then we get

$$\begin{aligned}\phi_{\widetilde{H}_t}(u) &= \mathbb{E}(\exp(iu\widetilde{H}_t)) \\ &= \mathbb{E}\left(\mathbb{E}\left[\exp(iu\widetilde{H}_t) \mid Y_t\right]\right) \\ &= \mathbb{E}\left(\exp(iu\widetilde{H}_{2,t}\ )\ \mathbb{E}\left[\exp(iu\widetilde{H}_{1,t}\ ) \mid Y_t\right]\right).\end{aligned}$$

Moreover

$$\begin{aligned}\mathbb{E}\left[\exp(iu\widetilde{H}_{1,t}) \mid Y_t\right] &= \phi_{\widetilde{H}_{1,t|Y_t}}(u) \qquad (8)\\ &= \phi_{N_x^{1/2}\left(\widehat{F}_t^{-1}(\widehat{G}_t(x))-F_t^{-1}(\widehat{G}_t(x))\right)|Y_t}\left(\left(N_y/(N_x+N_y)\right)^{1/2}u\right)\end{aligned}$$

From (7) it follows that, $\forall v \in \mathbb{R}$,

$$\phi_{N_x^{1/2}\left(\widehat{F}_t^{-1}(\widehat{G}_t(x))-F_t^{-1}(\widehat{G}_t(x))\right)|Y_t}(v) \longrightarrow \exp\left(-\frac{1}{2}v^2\widehat{\sigma}_t^2(x)\right), \qquad (9)$$

as $N_x \to \infty$, where

$$\widehat{\sigma}_t^2(x) = \frac{\widehat{G}_t(x)(1-\widehat{G}_t(x))}{f_t^2\left(F_t^{-1}(\widehat{G}_t(x))\right)}.$$

The convergence (5) yields $\widehat{G}_t(x) \xrightarrow{P} G_t(x)$, as $N_y \to \infty$, which implies, combined with (8)-(9), Assumption 1 and $h_t(x) = F_t^{-1}(G_t(x))$, that

$$\mathbb{E}\left[\exp(iu\widetilde{H}_{1,t}) \mid Y_t\right] \xrightarrow{d} \exp\left(-\frac{1}{2}(1-a)u^2\sigma_t^2(x)\right), \qquad (10)$$

as $N_x \to \infty$ and $N_y \to \infty$. Moreover we have

$$\exp(iu\widetilde{H}_{2,t}\ ) = \exp\left[iu\left(\frac{N_x}{N_x+N_y}\right)^{1/2} N_y^{1/2}\left(F_t^{-1}(\widehat{G}_t(x)) - F_t^{-1}(G_t(x))\right)\right].$$

Since the function $x \mapsto \exp(iux)$ is continuous, then the convergence (6) and Assumption 1 yield

$$\exp(iu\widetilde{H}_{2,t}\ ) \xrightarrow{d} \exp(iua^{1/2}H_{2,t}\ ), \qquad \text{as } N_x \to \infty, N_y \to \infty, \qquad (11)$$

where $H_{2,t}$ is centered Gaussian distributed with variance equal to $\sigma_t^2(x)$). From (10) and (11) it follows that, as $N_x \to \infty$ and $N_y \to \infty$,

$$\begin{aligned}&\exp(iu\widetilde{H}_{2,t}\ )\mathbb{E}\left[\exp(iu\widetilde{H}_{1,t}) \mid Y_t\right] \\ &\qquad \xrightarrow{d} \exp(iua^{1/2}H_{2,t}\ ) \qquad \exp\left(-\frac{1}{2}(1-a)u^2\sigma_t^2(x)\right). \qquad (12)\end{aligned}$$

Since $\mathbb{E}\left[\exp(iu\widetilde{H}_{1,t}) \mid Y_t\right]$ and $\exp(iu\widetilde{H}_{2,t})$ are bounded almost surely, it follows from (12) that

$$\begin{aligned}
\phi_{\widetilde{H}_t}(u) &= \mathbb{E}\left(\exp(iu\widetilde{H}_{2,t}\,)\mathbb{E}\left[\exp(iu\widetilde{H}_{1,t}) \mid Y_t\right]\right) \\
&\to \mathbb{E}\left(\exp\left(iua^{1/2}H_{2,t}\right)\exp\left(-\frac{1}{2}(1-a)u^2\sigma_t^2(x)\right)\right), \text{ as } N_x \to \infty, N_y \to \infty \\
&= \exp\left(-\frac{1}{2}au^2\sigma_t^2(x)\right)\exp\left(-\frac{1}{2}(1-a)u^2\sigma_t^2(x)\right) \\
&= \exp\left(-\frac{1}{2}u^2\sigma_t^2(x)\right),
\end{aligned}$$

therefore the desired conclusion (4) holds. ■

Lemma 2.1 implies that

$$\widehat{h}_t(x) = h_t(x) + \sigma_{1,t}(x)\varepsilon_t + r_t, \tag{13}$$

where $\sigma_{1,t}^2(x)$ is given by (2), (ε_t) is a standard Gaussian white noise and the remainder term r_t is such that

$$r_t = O_P\left(\{(N_x + N_x)/N_xN_y\}^{1/2}\right). \tag{14}$$

Let $D = D[0,1]$ be the space of random functions that are right-continuous and have left limits, endowed with the Skorohod topology. The weak convergence of a sequence of random elements X_n in D to a random element X in D will be denoted by $X_n \Longrightarrow X$. Let

$$W_n(\tau) = \frac{1}{\sqrt{n}}\sum_{t=1}^{[n\tau]}\sigma_{1,t}(x)\varepsilon_t, \quad \tau \in [0,1]. \tag{15}$$

LEMMA 2.2 *Under Assumptions 1-3 we have*

$$W_n \implies \overline{\sigma}_2(x)\, W, \tag{16}$$

where W stands for the standard Brownian motion.

Proof Assumption 2 implies that

$$\begin{aligned}
\sigma_{1,t}^2(x) &\leq \frac{1}{\gamma_1^2}\frac{N_x + N_y}{N_xN_y} \\
&\leq C,
\end{aligned}$$

for some positive constant C and N_x and N_y large enough. Hence $\sigma_{1,t}^2(x)$ is a bounded deterministic sequence, therefore the weak convergence (16) follows from Theorem A.1 of [5]. ■

LEMMA 2.3 *Under the null H_0, as $n \to \infty$, $N_x \to \infty$ and $N_y \to \infty$ we have*

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{t=1}^{n}(\widehat{h}_t(x) - \bar{h}(x))^2 \quad \underrightarrow{d} \quad \overline{\sigma}_2^2(x). \tag{17}$$

Proof Under the null H_0: $h_t(x) = h(x)$ the equality (13) becomes

$$\widehat{h}_t(x) \quad = \quad h(x) + \sigma_{1,t}(x)\varepsilon_t + r_t.$$

Let $y_t = h(x) + \sigma_{1,t}(x)\varepsilon_t$, $\overline{y} = \sum_{t=1}^{n} y_t/n$, then by using the same argument as in Theorem 1 of [5] we obtain

$$\frac{1}{n}\sum_{t=1}^{n}(y_t - \overline{y})^2 \quad \underrightarrow{d} \quad \overline{\sigma}_2^2(x). \tag{18}$$

We have

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}(\widehat{h}_t(x) - \bar{h}(x))^2 \\
&\quad = \frac{1}{n}\sum_{t=1}^{n}(y_t - \overline{y})^2 \quad + \quad \frac{1}{n}\sum_{t=1}^{n}(r_t - \overline{r})^2 + 2\frac{1}{n}\sum_{t=1}^{n}(y_t - \overline{y})(r_t - \overline{r}),
\end{aligned} \tag{19}$$

where $\overline{r} = \sum_{t=1}^{n} r_t/n$. From (14) it follows that

$$\begin{aligned}
\overline{r} \quad &= \quad O_P(((N_x + N_y)/N_x N_y)^{1/2}) \\
&= \quad o_p(1), \text{ as } \; N_x \to \infty, N_y \to \infty,
\end{aligned}$$

which implies that

$$\frac{1}{n}\sum_{t=1}^{n}(r_t - \overline{r})^2 = o_p(1), \qquad \text{as } \; N_x \to \infty, N_y \to \infty. \tag{20}$$

By using the Cauchy Shwartz inequality, we have

$$\frac{1}{n}\sum_{t=1}^{n}(y_t - \overline{y})(r_t - \overline{r}) \quad \leq \quad \left(\frac{1}{n}\sum_{t=1}^{n}(y_t - \overline{y})^2\right)^{1/2}\left(\frac{1}{n}\sum_{t=1}^{n}(r_t - \overline{r})^2\right)^{1/2}.$$

Hence by using (18) and (20) we get

$$\frac{1}{n}\sum_{t=1}^{n}(y_t - \overline{y})(r_t - \overline{r}) = o_p(1), \qquad \text{as } \; N_x \to \infty, N_y \to \infty. \tag{21}$$

The desired conclusion (17) holds by combining (18)-(21).

■

Proof of Theorem 1 Under the null, the process $B_n(\tau, x)$ in (1) can be rewritten as

$$\begin{aligned} B_n(\tau,x) &= \frac{1}{\sqrt{n}\widehat{\sigma}_n}\left(\sum_{t=1}^{[n\tau]}\sigma_{1,t}(x)\varepsilon_t - \frac{[n\tau]}{n}\sum_{t=1}^{n}\sigma_{1,t}(x)\varepsilon_t\right) + R_n(\tau,x) \\ &= \frac{1}{\widehat{\sigma}_n}\left(W_n(\tau) - \frac{[n\tau]}{n}W_n(1)\right) + R_n(\tau,x), \end{aligned}$$

where the remainder term $R_n(\tau, x)$ is given by

$$R_n(\tau,x) = \frac{1}{\sqrt{n}\widehat{\sigma}_n}\left(\sum_{t=1}^{[n\tau]} r_t - \frac{[n\tau]}{n}\sum_{t=1}^{n} r_t\right).$$

Now observe that

$$\sum_{t=1}^{[n\tau]} r_t = O_P\left([n\tau]\left((N_X+N_Y)/N_XN_Y\right)^{1/2}\right),$$

which together with (17) implies that

$$\begin{aligned} R_n(\tau,x) &= O_P\left(\{n(N_x+N_y)/N_xN_y\}^{1/2}\right), \\ &= o_p(1) \text{ under assumption 4.} \end{aligned}$$

Hence

$$R_n(\tau,x) = \frac{1}{\widehat{\sigma}_n}\left(W_n(\tau) - \frac{[n\tau]}{n}W_n(1)\right) + o_p(1),$$

which combined with (16) and (17) yields

$$B_n(.,x) \implies B,$$

where $B(\tau) = W(\tau) - \tau W(1)$ is a Brownian bridge. Therefore

$$\sup_{1\leq\tau\leq 1}|B_n(\tau,x)| \xrightarrow{d} \sup_{1\leq\tau\leq 1}|B(\tau)|. \tag{22}$$

Let $F(\mathbb{R},\mathbb{R})$ be the space of square integrable functions endowed with the uniform norm $\|.\|_\infty$. For a given square integrable function w, the functional $\mathcal{G}_w$: $(F(\mathbb{R},\mathbb{R}), \|.\|_\infty) \to (\mathbb{R}, |.|)$ defined by

$$\mathcal{G}_w(g) = \int_{\mathbb{R}} w(x)g(x)dx,$$

is continuous. To obtain the convergence (3) it is sufficient to apply (22) and the continuous mapping theorem.

■

3 Empirical study

For simplicity we consider $N_x = N_y = N$. Data are generated from three models: first, Y_t is normally distributed with mean 0 and variance 1, and X_t is generated independently by the transformation $X_t = h_t(Z_t)$, where Z_t is another Gaussian process with mean 0 and variance 1. Second, Y_t is an autoregressive process of order 1 (AR1) with correlation coefficient equal to 0.5, and X_t is generated independently by the transformation $X_t = h_t(Z_t)$, where Z_t is another AR1 process. For the last model random variables are paired: Y_t are independent Gaussian variables with mean 0 and variance 1, and $X_t = h_t(Y_t)$, that is, the time transformation is on the random variables. It is clear that this implies the same transformation for the corresponding distributions.

Alternatives. The following five alternatives are considered

First alternative: A1
Change in the mean. $h_{1,t}(x) = \dfrac{2t^2}{1+t^2} + x.$

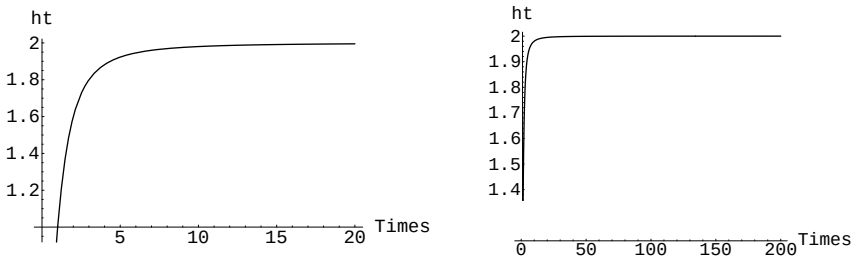
Second alternative: A2
Change in the variance. $h_{2,t}(x) = \dfrac{2t^2}{1+t^2}\, x.$

Third alternative: A3
Jump. $h_{3,t}(x) = x + 0.05t\mathbb{I}_{t<n/2} + 0.005(n-t)\mathbb{I}_{t\geq n/2}$,
where $\mathbb{I}_{t\geq n/2} = 1$ if $t \geq n/2$ and 0 otherwise.

Fourth alternative: A4
Smooth change in the mean. $h_{4,t}(x) = x + (1+\exp(-0.01(t-1)))^{-1}$

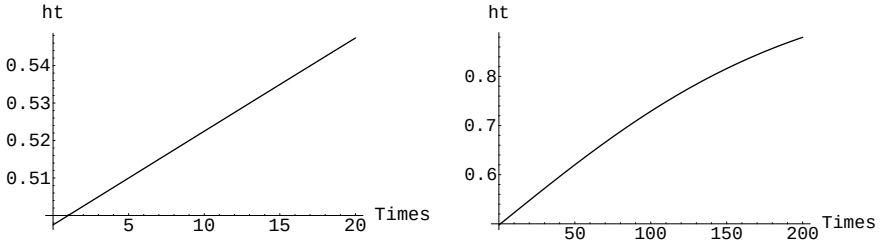
Fifth alternative: A5
Smooth change in the mean. $h_{5,t}(x) = x + (1+\exp(-0.05(t-1)))^{-1}$

All alternatives are smooth and are less rough than classical rupture on the mean or on the variance, except A3 which coincides with a jump on the mean. The first two alternatives A1-A2 tend quickly to the null model under H_0 when the length n increases. Figure 1 illustrates the proximity of h_t to a constant for large times length in the case of alternative A1. In opposition, alternatives A4-A5 are very smooth and converge slowly to the null model. Figure 2 illustrates this smooth convergence under alternative A4.



(a) Time length = 20 (b) Time length = 200

Figure 1: Representation under alternative A1 of $h_t = 2t^2/(1+t^2)$ for time length = 20 (a) and time length = 200 (b)



(a) Time length = 20 (b) Time length = 200

Figure 2: Representation under alternative A4 of $h_t = (1+\exp(-0.01(t-1)^2))^{-1}$ for time length = 20 (a) and time length = 200 (b)

Bootstrap procedure. To evaluate the power of our testing procedure we first consider a Monte Carlo statistic. Given M points $x_1, \cdots, x_M$ in $\mathcal{Y}$ we consider

$$S_M(w) = \frac{1}{M}\sum_{i=1}^{M} w(x_i)A(x_i), \tag{23}$$

where

$$A(x_i) \quad = \quad \max_{1\leq k\leq n} \left| \frac{1}{\widehat{\sigma}_n(x_i)\sqrt{n}} \left(\sum_{t=1}^{k} \widehat{h}_t(x_i) - k\widehat{h}(x_i) \right) \right|,$$

with

$$\begin{cases} \widehat{\sigma}_n^2(x) &= \dfrac{1}{n}\displaystyle\sum_{t=1}^{n}(\widehat{h}_t(x) - \bar{h}(x))^2 \\ \bar{h}(x) &= \dfrac{1}{n}\displaystyle\sum_{t=1}^{n}\widehat{h}_t(x). \end{cases}$$

The convergence of the statistic $S_M(w)$ is not guaranteed since the $A(x_i)$ are dependent. To carry out this problem, a bootstrap procedure is proposed. We construct a naive bootstrap statistic; that is, the test statistic $S_M(w)$ given in (23) is compared to the empirical bootstrapped distribution obtained from $({S_M}^{*b})_{b=1,\cdots,B}$, with S_M^{*b} constructed from the bootstraped sample drawn randomly with replacement and satisfying the size equalities $N_x^* = N_x$ and $N_y^* = N_y$. We fix w as a constant. Note that if X and Y are paired, the bootstrap procedure consists in drawing randomly with replacement N pairs (X, Y) from the data. We fix $B = 200$ bootstrap replications.

Powers. For each alternative, the test statistic is computed, based on sample sizes $N = 50, 100$, for a theoretical level $\alpha = 5\%$. The lengths of time's intervals are $n = 20, 100$ and 200; that is, the function h_t is observed N times for each t varying in $[0; 20]$, or $[0; 100]$, or $[0; 200]$, with a step equal to one. The empirical power of the test is defined as the percentage of rejection of the null hypothesis over 10000 replications of the test statistic under the alternative.

Figure 3 presents empirical powers of the bootstrap test for all alternatives, in the case where X_t are independent standard Gaussian variables. Solid lines and dotted lines correspond to $N = 50$ and 100 respectively. It can be observed that the power decreases with the length for alternatives A1 and A2. It is in accordance with the previous remark: h_t is close to the null hypothesis for relatively large values of n. Then passing from a time length equal 20 to a time length equal to 200 corresponds to adding variables with nearly constant transformation in distribution (see Figure 1).

Alternatives A4-A5 have similar behaviors, with a power increasing with n. It can be explained by the very slow convergence to the null model. Here, passing from a time length equal 20 to a time length equal to 200 corresponds to adding new observations with a time depending transformation (see Figure 2).

It is also observed that power associated to alternative A3 increases with n.

In Figure 4 empirical powers are presented in the case where Y_t follows an AR1 process with a correlation coefficient equal to 0.5. Here powers are slightly better and more stable with respect to the length. This is due to the correlation inducing more stability of the process Y_t and permitting a better estimation of h_t.

Figure 5 presents results in the case of paired data, with Y_t normally distributed. Powers are good, due to the fact that transformations occur not randomly since we have considered $X_t = h_t(Y_t)$. Then h_t can be efficiency estimated and its variations are well detected.

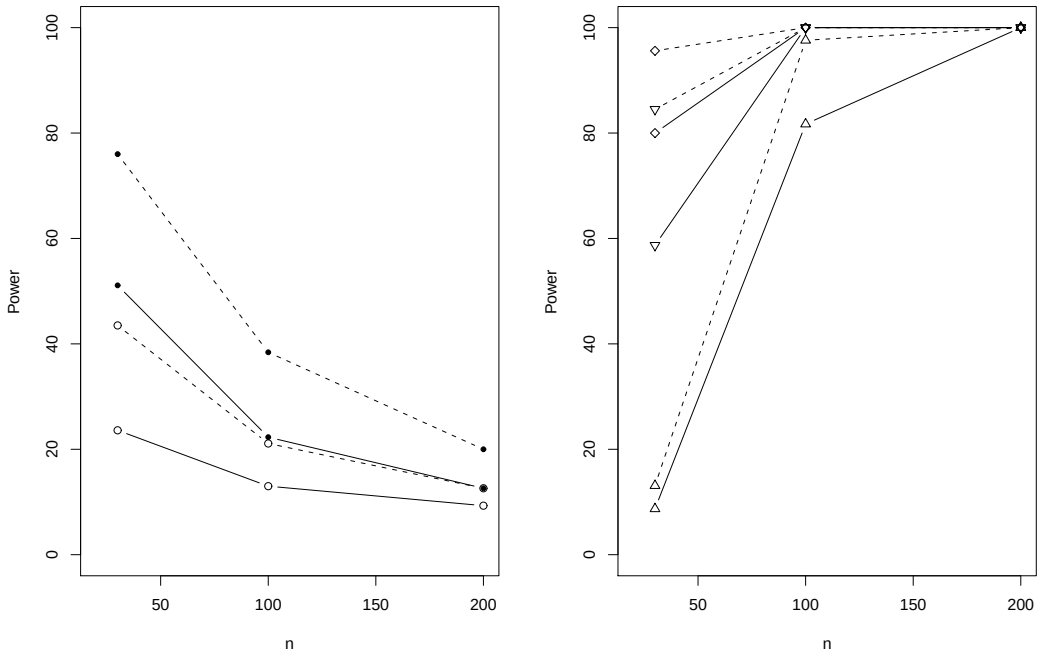


Figure 3: Empirical powers for alternatives A1 (•) and A2 (○) on the left, A3 (◇), A4 (△) and A5 (▽) on the right, with X_t distributed as $\mathcal{N}(0,1)$. Solid lines correspond to $N = 50$ and dotted lines correspond to $N = 100$. The lengths of time's intervals are $n = 20, 100, 200$

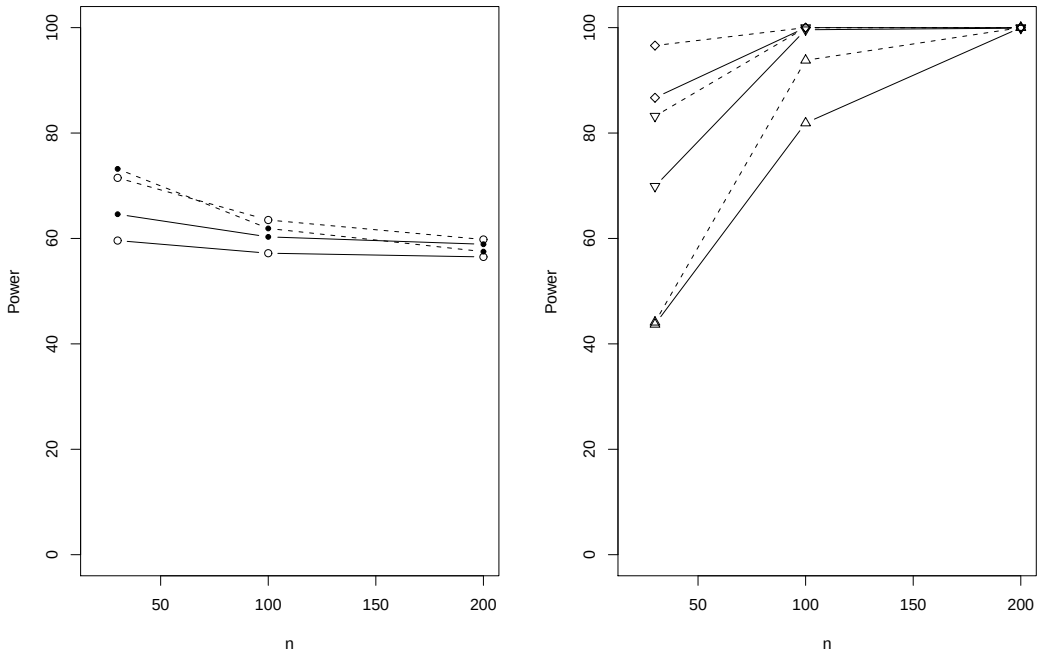


Figure 4: Empirical powers for alternatives A1 (•) and A2 (○) on the left, A3 (◇), A4 (△) and A5 (▽) on the right, with X_t following an AR1 process with correlation 0.1. Solid lines correspond to $N = 50$ and dotted lines correspond to $N = 100$. The lengths of time's intervals are $n = 20, 100, 200$

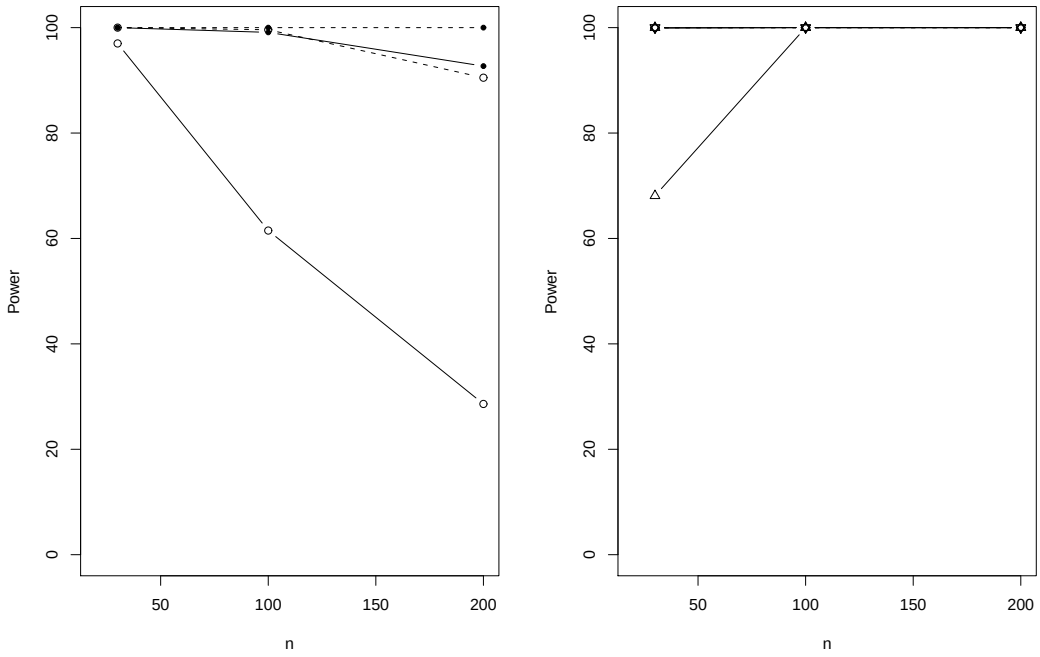


Figure 5: Empirical powers for alternatives A1 (•) and A2 (○) on the left, A3 (◇), A4 (△) and A5 (▽) on the right, with X_t and Y_t paired. Solid lines correspond to $N = 50$ and dotted lines correspond to $N = 100$. The lengths of time's intervals are $n = 20, 100, 200$

4 Concluding remarks

The proposed method concerns the comparison of two processes when panel data are available. The test permits to detect a change in the relation between the two process distributions. Therefore it can detect a change in a higher moments (not only in the mean and/or in the variance as almost tests do in this framework). The asymptotic distribution of the proposed statistic was derived under the null of no change in the relation between the two process distributions.

The Monte Carlo simulations show that our test performs well in finite sample and has a good power against either abrupt or smooth changes. It is also valid for paired processes and then it can be used to detect a change in h_t in the relation $X_t = h_t(Y_t)$ (see the paired case in our simulations). The test can also be used as a first step permitting to legitimate estimation and interpretation of a constant transformation h between two panel data, as for instance in a

medical follow-up study.

A direction for future research is to consider a d-sample comparison of distributions, for $d > 2$, in the way of [3, 2]. Another direction should consider multivariate distributions.

References

[1] Balakrishnan, N., Xingqiu Zhao b. (2010). A nonparametric test for the equality of counting processes with panel count data *Computational Statistics & Data Analysis* 54, 135–142.

[2] Balakrishnan, N., Xingqiu Zhao, b. (2009). New multi-sample nonparametric tests for panel count data. *The Annals of Statistics* 37, 1112–1149.

[3] Balakrishnan, N., Xingqiu Zhao, b. (2008). A class of multi-sample nonparametric tests for panel count data. *Ann. Instit. Statist. Math.* 60, 151–171.

[4] Billingslley, P. (1968). *Convergence of probability measures*, Wiley, New York.

[5] Boutahar, M. (2009). Testing for change in the mean of heteroskedastic time series. http://fr.arxiv.org/abs/1102.5431.

[6] Davidson, J. (1994). *Stochastic Limit Theory.* Oxford: Oxford University Press.

[7] Galeano, P., Peña, D. (2007). Covariance changes detection in multivariate time series *Journal of Statistical Planning and Inference* 137, 194 – 211.

[8] Gombay, E. (2008). Change detection in autoregressive time series, *J. Multivariate Anal.* 99, 451-464.

[9] Gupta A.K., Tang J. (1984). Distribution of likelihood ratio statistic for testing equality of covariance matrices of multivariate Gaussian models *Biometrika* 71, 555–559.

[10] Hawkins, D.M. (1977). Testing a sequence of observations for a shift in location, *J. Amer. Statist. Assoc.* 72, 180–186.

[11] James, B., James, K., Siegmund, D. (1987). Tests for a change point, *Biometrika* 74, 71– 83.

[12] Panaretos, V.M., Kraus, D. & Maddocks, J.H. (2009). Second-Order Comparison of Gaussian Random Functions and the Geometry of DNA Minicircles. *Journal of the American Statistical Association* 490, 670–682.

[13] Sen, A., Srivastava, M.S. (1975). On tests for detecting change in mean, *Ann. of Statist.* 3,1 98– 108.

[14] Srivastava, M.S., Worsley, K.J. (1986). Likelihood ratio tests for a change in the multivariate normal mean, *J. Amer. Statist. Assoc.* 81, 199–204.

[15] Worsley, K.J. (1979). On the likelihood ratio test for a shift in locations of normal populations, *J. Amer. Statist. Assoc.* 74, 365– 367.