Annals of the Institute of Statistical Mathematics (2021) 73:1063–1087
https://doi.org/10.1007/s10463-021-00786-3

Distribution-free testing in linear and parametric regression

Estate V. Khmaladze[1]

Received: 30 January 2020 / Revised: 30 November 2020 / Accepted: 4 December 2020 /
Published online: 8 March 2021

Abstract
Recently, a distribution-free approach for testing parametric hypotheses based on unitary transformations has been suggested in Khmaladze (Ann Stat 41:2979–2993, 2013, Bernoulli 22:563–588, 2016) and further studied in Nguyen (Metrika 80:153–170, 2017) and Roberts (Stat Probab Lett 150:47–53, 2019). In this paper, we show that the transformation takes very simple form in distribution-free testing of linear regression. Then, we extend it to the general parametric regression with vector-valued covariates.

Keywords Regression empirical process · Unitary operators · Distribution-free residuals · Linear regression · Optimal transport

1 Introduction: an illustrative example with linear regression

The situation we consider in this paper is that of the classical parametric regression: given a sequence of pairs of random variables $(X_i, Y_i)_{i=1}^n$, where Y_i is the response variable, while X_i is the explanatory variable, or covariate, of this Y_i, consider regression of Y_i on X_i,

$$Y_i = m(X_i) + \epsilon_i.$$

We assume that, given covariates $(X_i)_{i=1}^n$, the errors $(\epsilon_i)_{i=1}^n$ are i.i.d. and have expected value zero and finite variance—for the sake of simplicity, we assume this variance equal 1.

We are interested in the classical problem of testing that the regression function $m(x)$ belongs to a specified parametric family of functions $(m(x,\theta), \theta \in \Theta)$, which depend on a finite-dimensional parameter θ and which satisfy more or less usual regularity assumptions as functions of this θ.

Estate V. Khmaladze
Estate.Khmaladze@vuw.ac.nz

[1] Victoria University of Wellington, PO Box 600, Wellington, New Zealand

Our aim is to describe a new method to build asymptotically distribution-free theory for testing such hypotheses. More specifically, we will construct asymptotically distribution-free version of the regression empirical process, so that functionals from this process, used as test statistics, will be asymptotically distribution-free. The core of the method consists of the application of unitary operators as described more or less recently in Khmaladze (2013), Khmaladze (2016), Khmaladze (2020) and studied in Roberts (2019) and Nguyen (2017).

Earlier, asymptotically distribution-free transformation of regression empirical process was suggested in Khmaladze and Koul (2004), see also for quantile regression Koenker (2005). For d-dimensional covariates, the limit distribution of the transformed process was that of standard Brownian motion on $[0,1]^d$. In this paper, the transformed process will converge to a standard projection of the standard Brownian motion on $[0,1]^d$, and the transformation will take surprisingly simple form, convenient in everyday practice. In the case of linear regression, it should be called elementary. As in Khmaladze and Koul (2004), this transformation is connected with no loss of statistical information.

The shortest way to show how the method works is to consider the most simple linear regression model. That is, in

$$Y_i = X_i\theta + \epsilon_i,\ i = 1, \dots, n, \text{ or in vector form, } \quad Y = X\theta + \epsilon, \tag{1}$$

the covariates X_i, and the coefficient θ are one-dimensional. On probabilistic nature of the covariates $(X_i)_{i=1}^n$, we will make, practically, no assumptions. We only will use their empirical distribution function

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{(X_i \le x)}$$

and assume that, as number of observed pairs n increases, it weakly converges to some limiting distribution F—an assumption of ergodic nature. Whenever we use time transformation $t = F(x)$, we will also assume that F is continuous. All expectations below will be conditional expectations given the vector of numbers $(X_i)_{i=1}^n$.

Consider estimated errors, or residuals,

$$\hat{\epsilon} = Y - X\hat{\theta},$$

where $\hat{\theta} = \langle Y, X\rangle / \langle X, X\rangle$ denotes the least square estimator of θ. It is convenient to re-write $\hat{\epsilon}$ as

$$\hat{\epsilon} = Y - z\langle Y, z\rangle = \epsilon - z\langle \epsilon, z\rangle, \quad \text{where} \quad z = X/\langle X, X\rangle^{1/2},$$

which represents the vector of residuals as projection of ϵ orthogonal to the vector of normalised covariates z.

The natural object to base a goodness-of-fit test upon is given by the partial sums process (see, for example, Khmaladze and Koul 2004 and Stute 1997)

$$\hat{w}_n(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{\epsilon}_i \mathbb{I}_{(X_i \le x)}.$$

However, the distribution of the vector $\hat{\epsilon}$ depends on covariates: its covariance matrix has the form

$$E\hat{\epsilon}\,\hat{\epsilon}^T = I - zz^T.$$

As to the limit in distribution for the process $\hat{w}_n$, it is a projection of some Brownian motion, but it is not the Brownian bridge. Its distribution remains dependent on behaviour of the covariates. The limit distribution of omnibus statistics based on this process, and in particular, its supremum, will not be easy to calculate.

However, consider new residuals obtained from $\hat{\epsilon}$ by unitary transformation

$$U_{a,b} = I - \frac{\langle a-b, \cdot\rangle}{1 - \langle a, b\rangle}(a-b)$$

with n-dimensional vectors a and b of unit norm: $\|a\| = \|b\| = 1$. If $a = b$, we take $U_{a,b} = I$. This operator is unitary, it maps a into b and b into a, and it maps any vector c, orthogonal to a and b, to itself (see, for example, Khmaladze 2013, Sec. 2). Now choose $a = z$ and choose b equal $r = (1, \ldots, 1)^T/\sqrt{n}$, the vector not depending on covariates at all. Since the vector of residuals $\hat{\epsilon}$ is orthogonal to the vector z, we obtain:

$$\hat{e} = U_{r,z}\hat{\epsilon} = \hat{\epsilon} - \frac{\langle \hat{\epsilon}, r\rangle}{1 - \langle z, r\rangle}(r - z).$$

These new residuals have covariance matrix

$$E\hat{e}\hat{e}^T = I - rr^T.$$

This would be the covariance matrix of the residuals in the problem of testing

$$Y_i = \theta + \epsilon_i, \quad i = 1, 2, \ldots, n, \tag{2}$$

which is completely free from covariates. Yet, the transformation of $\hat{\epsilon}$ to $\hat{e}$ is one-to-one, and therefore, $\hat{e}$ contain the same "statistical information", whichever way we measure it, as $\hat{\epsilon}$. One could say that the problem of testing linear regression (1) and testing (2) is the *same* problem.

The partial sum process based on the new covariates,

$$\hat{w}_{n,e}(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{e}_i \mathbb{I}_{(X_i \le x)},$$

will converge in distribution, with time transformation $t = F(x)$, to standard Brownian bridge. Therefore, limit distribution for all classical statistics will be free from covariates and known (cf. Fig. 1).

Asymptotically distribution-free tests, even if only for the case of linear regression, have been of main interest from long ago. To achieve this distribution-freeness,

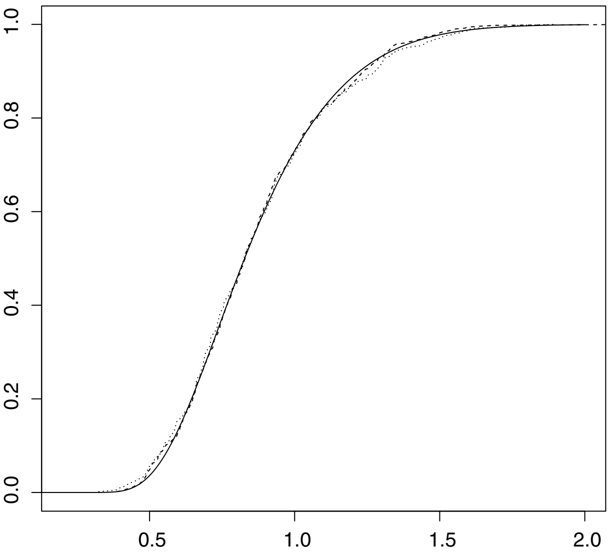

Fig. 1 The smooth line is Kolmogorov distribution function. The two other ones are simulated distributions of $\max_x |\hat{w}_{n,e}(x)|$ for two entirely different behaviour of covariates. In one case X_i-s have uniform distribution on [0, 2] while in the other they have Gaussian distribution $N(1, 2)$. 200 replications of samples of size $n = 200$.

different forms of residuals have been suggested, various decompositions of z, especially when covariates X_i are multidimensional, have been studied and approximations for quadratic forms from $\hat{\epsilon}$ have been developed. Assumption of normality, arbitrary as it is in many cases, has been made more or less casually. If one is allowed somewhat free speech, one could say that a mathematical lace has been created. Good source for this material is the book Cook and Weisberg (1982). In dry residue, only the Chi-square tests have been obtained. Distribution-free forms of other classical statistics have never been considered and constructed. We refer to McCullagh and Nelder (2008) for much of the existing theory for linear models. The most recent review on goodness-of-fit problems in regression which we know of is Gonzalez-Manteiga and Crujeiras (2013).

Note that the initial regression process of this paper, not yet asymptotically distribution-free, is different from what was used in previous work, including relatively recent ones. Although partial sum processes, like $\hat{w}_n$, form one of the main objects of asymptotic theory, it is often that a different form of such processes is considered, one simple example of which would be

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \bar{X}_n) \mathbb{1}_{(\hat{\epsilon}_i \le x)}, \tag{3}$$

(see more sophisticated form of the weight function in recent paper Chown and Müller 2018). Here the scanning over the values of the residuals is used. This is very natural way of scanning when the statistical problems considered pertain to distribution of errors. An example, studied in well-known papers Dette and Munk (1998), Dette and Hetzler (2009), Dette et al. (2007) and loc.cit. Chown and Müller (2018), is the problem of testing heterogeneity of errors. The same scanning is basically unavoidable in study of distribution of i.i.d. errors, cf. Koul et al. (2017), and in analysis of the distribution of innovations in autoregression models, see Müller et al. (2009).

In our current situation of testing the form of regression function, it is a natural wish to see, in the case there is a deviation from the model, for what region of values of the covariate the deviation takes place, and scanning in X_i-s will allow this. Even in the simple case when the covariate is just discrete time, taking values $1, 2, \dots, n$, it would be strange not to examine the sequence $\hat{\epsilon}_1, \dots, \hat{\epsilon}_n$, in this time, but instead look on the order statistics based on them, which scanning as in (3) would imply. These considerations motivate the form of the regression process $\hat{w}_n$ and $\hat{w}_{n,e}$.

To make the illustrative example of this section more of immediate practical use and to explain better the asymptotic behaviour of the regression empirical process, in Sect. 2 we consider the general form of one-dimensional linear regression. Then in Sect. 3 we consider general parametric regression. In this case, the time transformation, considered in (iii) of Proposition 2 again leads to distribution-freeness if F is continuous. If F is discrete, then the method suggested in Khmaladze (2013), Sec. 2, can be easily used. In Sect. 4, we consider multidimensional X_is. Transformation of $\hat{\epsilon}$ to $\hat{e}$ will not change, but to standardise distribution of regressors one could use normalisation by $\hat{f}_n^{1/2}$, where $\hat{f}_n$ is an estimator of the density of F, cf., e.g., Einmahl and Khmaladze (2001), Can et al. (2020). Here, however, we consider an approach borrowed from the theory of optimal transportation, or Monge–Kantorovich transportation problem, see, for example, Villani (2009). Very interesting probabilistic/statistical applications of this theory have been recently given in del Barrio et al. (2018) and de Valk and Segers (2018).

2 General linear regression on $\mathbb{R}$

Consider the standard linear regression on the real line,

$$Y_i = \theta_0 + \theta_1 X_i + \epsilon_i, \; i = 1, \dots, n, \text{ or } \quad Y = \theta_0 \mathbf{1} + \theta_1 X + \epsilon. \tag{4}$$

The $\mathbf{1}$ here denotes a vector with all coordinates equal to the number 1. Instead of (4) consider its slightly modified and more convenient form

$$\begin{aligned} &Y_i = \theta_0 + \theta_1 (X_i - \bar{X}) + \epsilon_i, \; i = 1, \dots, n, \quad \text{or in vector form,} \\ &Y = \theta_0 \mathbf{1} + \theta_1 (X - \bar{X}\mathbf{1}) + \epsilon. \end{aligned} \tag{5}$$

The least square estimations of θ_0 and θ_1 are

$$\hat{\theta}_0 = \frac{1}{n} \sum_{j=1}^{n} Y_j \quad \text{and} \quad \hat{\theta}_1 = \frac{1}{\sum_{j=1}^{n} (X_j - \bar{X})^2} \sum_{i=1}^{n} Y_j (X_j - \bar{X}).$$

Using again notation $r = \mathbf{1}/\sqrt{n}$ and notation

$$\tilde{z} = \frac{1}{\sqrt{\sum_{j=1}^{n}(X_j - \bar{X})^2}}(X - \bar{X}),$$

for normalised vector of centred covariates, one can write the residuals as

$$\hat{\epsilon} = Y - \hat{\theta}_0\mathbf{1} - \hat{\theta}_1(X - \bar{X}\mathbf{1})$$

or in more succinct form

$$\hat{\epsilon} = Y - \langle Y, r\rangle r - \langle Y, \tilde{z}\rangle \tilde{z}.$$

Substitution of the linear regression model (5) for Y produces representation of the vector of residuals $\hat{\epsilon}$ through the vector of errors ϵ:

$$\hat{\epsilon} = \epsilon - \langle \epsilon, r\rangle r - \langle \epsilon, \tilde{z}\rangle \tilde{z}. \tag{6}$$

This represents $\hat{\epsilon}$ as projection of ϵ orthogonal to r and $\tilde{z}$.

From this, it follows that the covariance matrix of $\hat{\epsilon}$ is

$$E\hat{\epsilon}\hat{\epsilon}^T = I - rr^T - \tilde{z}\tilde{z}^T,$$

and thus it still depends on the values of the covariates. The limit distribution of the regression process with these residuals,

$$\hat{w}_n(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\epsilon}_i \mathbb{1}_{(X_i \le x)},$$

will therefore have limit distribution which depends on $\tilde{z}$.

It is possible to say more about the geometric structure of $\hat{w}_n$ and its limiting process, and namely that the limiting process will be a double projection of Brownian motion orthogonal to the functions $F(x)$ and

$$H(x) = \int^{x} h(y)\mathrm{d}F(y), \quad \text{with} \quad h(x) = \frac{x - \int y\mathrm{d}F(y)}{\sqrt{\int (z - \int y\mathrm{d}F(y))^2\mathrm{d}F(z)}}.$$

Here one can think of h as a continuous time version of $\tilde{z}$: the latter can be calculated by the same formula as h, with F replaced by F_n and x restricted to points X_i-s.

To show this structure of $\hat{w}_n$ denote $\mathbb{1}_x$, the vector with coordinates $(\mathbb{1}_{(X_i \le x)})_{i=1}^{n}$. Then, we can write

$$\hat{w}_n(x) = \frac{1}{\sqrt{n}}\langle \hat{\epsilon}, \mathbb{1}_x\rangle = \frac{1}{\sqrt{n}}\big[\langle \epsilon, \mathbb{1}_x\rangle - \langle \epsilon, r\rangle\langle r, \mathbb{1}_x\rangle - \langle \epsilon, \tilde{z}\rangle\langle \tilde{z}, \mathbb{1}_x\rangle\big].$$

For the first term on the right-hand side, considered as a process in x and denoted $w_n(x)$, we see that

$$w_n(x) = \frac{1}{\sqrt{n}}\langle \epsilon, \mathbb{1}_x \rangle = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i \mathbb{1}_{(X_i \leq x)} \tag{7}$$

is the process of partial sums of i.i.d. random variables and $Ew_n^2(x) = F_n(x)$, while $F_n \to F$. Therefore, w_n converges in distribution to Brownian motion in time F, i.e. $Ew_F^2(x) = F(x)$. Now consider the second term:

$$\frac{1}{\sqrt{n}}\langle \epsilon, r \rangle \langle r, \mathbb{1}_x \rangle = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \epsilon_j \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(X_i \leq x)} = w_n(\infty) F_n(x).$$

The third term produces the following expression:

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{j=1}^{n} \epsilon_j (X_j - \bar{X}) \frac{1}{\sum_{j=1}^{n} (X_j - \bar{X})^2} \sum_{i=1}^{n} (X_i - \bar{X}) \mathbb{1}_{(X_i \leq x)} \\
&= \int (y - \bar{X}) w_n(\mathrm{d}y) \frac{1}{\int (y - \bar{X})^2 \mathrm{d}F_n(y)} \int^{x} (y - \bar{X}) \mathrm{d}F_n(y) \\
&= \int h_n(y) w_n(\mathrm{d}y) \int^{x} h_n(y) \mathrm{d}F_n(y),
\end{aligned}$$

where

$$h_n(x) = \frac{x - \bar{X}}{\sqrt{\int (y - \bar{X})^2 \mathrm{d}F_n(y)}}.$$

This function, obviously, has unit $L_2(F_n)$-norm and is orthogonal to functions *const*. Overall, we see that

$$\hat{w}_n(x) = w_n(x) - w_n(\infty) F_n(x) - \int h_n(y) w_n(\mathrm{d}y) \int^{x} h_n(y) \mathrm{d}F_n(y), \tag{8}$$

and the right-hand side of (8) is the orthogonal projector of w_n, which annihilates F_n and H_n. As the consequence of this, if $\int y^2 \mathrm{d}F(y) < \infty$, then $\hat{w}_n$ will converge to the corresponding projection of the Brownian motion w_F.

What we propose now is, again, to replace the residuals $\hat{\epsilon}$ by another residuals, $\hat{e}$, constructed as their unitary transformation. As a preliminary step, assume that the covariates are listed in increasing order, $X_1 < X_2 < \cdots < X_n$. One can assume this without loss of generality: even if it will entail re-shuffling of our initial pairs of observations, the probability measure we work under will not change, because the re-shuffled errors will still be independent from permuted $(X_i)_{i=1}^n$ and will still form an i.i.d. sequence.

Now introduce another vector $\tilde{r}$, different from $\tilde{z}$, which also has unit norm and is orthogonal to r. Define

$$\hat{e} = U_{\tilde{z},\tilde{r}}\hat{\epsilon} = \hat{\epsilon} - \frac{\langle \hat{\epsilon}, \tilde{r} - \tilde{z}\rangle}{1 - \langle \tilde{z}, \tilde{r}\rangle}(\tilde{r} - \tilde{z}) = \hat{\epsilon} - \frac{\langle \hat{\epsilon}, \tilde{r}\rangle}{1 - \langle \tilde{z}, \tilde{r}\rangle}(\tilde{r} - \tilde{z}).$$

The second equality here is true because the vector $\hat{\epsilon}$ is orthogonal to the vector $\tilde{z}$, see (6). Thus, calculation of new residuals is as simple as in the previous case of (1).

Let us summarise properties of $\hat{e}$ in the following proposition. In this, for transition to the limit when $n \to \infty$, it is natural to assume that $\tilde{r}_i$ can be represented through some piece-wise continuous function $\tilde{r}(t)$ on [0, 1]:

$$\tilde{r}_i = \frac{1}{\sqrt{n}}\tilde{r}(\frac{i}{n}), \tag{9}$$

in which case we have convergence

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{nt}\tilde{r}_i = \frac{1}{n}\sum_{i=1}^{nt}\tilde{r}(\frac{i}{n}) \to \int_0^t \tilde{r}(s)\mathrm{d}s = Q(t)$$

and

$$\sum_{i=1}^{nt}\tilde{r}_i^2 = \frac{1}{n}\sum_{i=1}^{nt}\tilde{r}^2(\frac{i}{n}) \to \int_0^t \tilde{r}^2(s)\mathrm{d}s.$$

Orthogonality of the vector $\tilde{r}$ to the vector r implies orthogonality of the function $\tilde{r}(t)$ to functions equal constant, or $Q(1) = 0$. For example, $\tilde{r}$ can be chosen as

$$\tilde{r}_i = \sqrt{\frac{12}{n}}\Big[\frac{i}{n} - \frac{n+1}{2n}\Big]. \tag{10}$$

Proposition 1

(i) *Covariance matrix of* $\hat{e}$ *is*

$$E\hat{e}\hat{e}^T = I - rr^T - \tilde{r}\tilde{r}^T$$

and therefore does not incorporate covariates X as soon as $\tilde{r}$ *does not incorporate X.*

(ii) *If* (9) *is true, then the regression empirical process based on* $\hat{e}$,

$$\hat{w}_{n,e}(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{e}_i\mathbb{1}_{(X_i \le x)},$$

has the covariance function

$$E\hat{w}_{n,e}(x)\hat{w}_{n,e}(y) = F_n(\min(x,y)) - F_n(x)F_n(y) - Q_n(F_n(x))Q_n(F_n(y)),$$

where $Q_n(t) = \sum_{i=1}^{nt}\tilde{r}(\frac{i}{n})/n$. *In the case of* (10)

$$Q_n(F_n(x)) \sim -\sqrt{3}F_n(x)(1 - F_n(x)),\ n \to \infty.$$

(iii) *The process* $\hat{w}_{n,e}$*, with change of time* $t = F(x)$*, converges in distribution to projection of standard Brownian motion on* $[0, 1]$ *orthogonal to functions* 1 *and* $\tilde{r}$*.*

The main step in the proof of (i) is to express $\hat{e}$ through ϵ:

$$\begin{aligned} U_{\tilde{z},\tilde{r}}\hat{\epsilon} &= U_{\tilde{z},\tilde{r}}\epsilon - \langle \epsilon, r\rangle U_{\tilde{z},\tilde{r}}r - \langle \epsilon, \tilde{z}\rangle U_{\tilde{z},\tilde{r}}\tilde{z} \\ &= U_{\tilde{z},\tilde{r}}\epsilon - \langle \epsilon, r\rangle r - \langle \epsilon, \tilde{z}\rangle \tilde{r}, \end{aligned}$$

where the second equality is correct because $r \perp \tilde{z}, \tilde{r}$ and $U_{\tilde{z},\tilde{r}}\tilde{z} = \tilde{r}$ by the definition of $U_{\tilde{z},\tilde{r}}$. However, here the vector $e = U_{\tilde{z},\tilde{r}}\epsilon$ is the vector with independent standard normal coordinates,

$$EU_{\tilde{z},\tilde{r}}\epsilon\, \epsilon^T U_{\tilde{z},\tilde{r}} = U_{\tilde{z},\tilde{r}}U_{\tilde{z},\tilde{r}} = I,$$

because ϵ has independent standard normal coordinates and $U_{\tilde{z},\tilde{r}}$ is unitary and self-adjoint. At the same time,

$$\langle \epsilon, r\rangle = \langle U_{\tilde{z},\tilde{r}}\epsilon, U_{\tilde{z},\tilde{r}}r\rangle = \langle e, r\rangle$$

and

$$\langle \epsilon, \tilde{z}\rangle = \langle U_{\tilde{z},\tilde{r}}\epsilon, U_{\tilde{z},\tilde{r}}\tilde{z}\rangle = \langle e, \tilde{r}\rangle.$$

Therefore

$$\hat{e} = U_{\tilde{z},\tilde{r}}\hat{\epsilon} = e - \langle e, r\rangle r - \langle e, \tilde{r}\rangle \tilde{r},$$

which represents it as projection of e orthogonal to r and $\tilde{r}$ with covariance matrix given in (i).

To show (ii), use vector notation for $\hat{w}_{n,e}$:

$$E\hat{w}_{n,e}(x)\hat{w}_{n,e}(y) = \frac{1}{n}E\langle \square_x, \hat{e}\rangle\langle \hat{e}, \square_y\rangle = \frac{1}{n}\square_x^T(I - rr^T - \tilde{r}\tilde{r}^T)\square_y.$$

Opening the brackets in the last expression, one can find that

$$\frac{1}{n}\langle \square_x, \square_y\rangle = F_n(\min(x, y)) \quad \text{and} \quad \frac{1}{n}\langle \square_x, r\rangle\langle \square_y, r\rangle = F_n(x)F_n(y),$$

while

$$\begin{aligned} \frac{1}{n}\langle \square_x, \tilde{r}\rangle\langle \square_y, \tilde{r}\rangle &= \frac{1}{n}\sum_{i=1}^{n} \tilde{r}(\frac{i}{n})\square_{(X_i \le x)}\frac{1}{n}\sum_{i=1}^{n} \tilde{r}(\frac{i}{n})\square_{(X_i \le y)} \\ &= \frac{1}{n}\sum_{i=1}^{nF_n(x)} \tilde{r}(\frac{i}{n})\,\frac{1}{n}\sum_{i=1}^{nF_n(y)} \tilde{r}(\frac{i}{n}) = Q_n(F_n(x))Q_n(F_n(y)), \end{aligned}$$

which proves (ii).

The statement (iii) follows if we note that the covariance function of $\hat{w}_{n,e}(x)$ in time $t = F(x)$ converges to $\min(t, s) - ts - Q(t)Q(s)$ and that orthogonality of function $\tilde{r}(\cdot)$ to the function identically equal 1 makes the last expression the covariance of the Gaussian process

$$w(t) - tw(1) - Q(t) \int_0^1 \tilde{r}(s) w(\mathrm{d}s),$$

which indeed is the projection described in (iii). □

In both regression models (1) and (5), but let us speak first about model (5), the process $\hat{w}_n$ turns out to be a two-dimensional projection of a Brownian motion. When the values of the covariates change, this projection will change. However, it is geometrically clear that it should be possible to rotate one projection into another, and this another into still another one, thus creating a class of equivalent projections—those which can be mapped into each other. Then, one can choose a single representative in each equivalence class, call it standard and rotate any other projection into this standard one. What was done in this and the previous section was that we selected two standard projections and constructed the rotation of the other ones into these two.

The practical usefulness of this approach depends on how simple the rotation is. For us, the transformations of $\hat{\epsilon}$ into $\hat{e}$ look very simple.

Finally, note that model (5) includes two estimated parameters, while model (1)—only one. However, since the vector r is already "standard", independent from covariates, there is no need to "rotate" it to any other vector. Therefore, in both cases one-dimensional rotation is sufficient. The situation when one needs to rotate several vectors at once, as well as general form of parametric regression, will be considered in the next Sect. 3.

3 General parametric regression

Now consider testing regression model

$$Y_i = m_\theta(X_i) + \epsilon_i,\ i = 1, \ldots, , n, \quad \text{or in vector form, } Y = m_\theta(X) + \epsilon, \tag{11}$$

where $m_\theta(X)$ denotes a vector with coordinates $(m_\theta(X_i))_{i=1}^n$, and m_θ is regression function, depending on d-dimensional parameter θ. We will assume some regularity of $m_\theta(X_i)$ with respect to θ, namely that $m_\theta(X_i)$ is continuously differentiable in θ. Our first aim is to demonstrate that in general situation the regression empirical process $\hat{w}_n$ asymptotically remains a projection of a Brownian process.

3.1 Regression empirical process with estimated parameters

The main steps in this subsection may, for some readers, sound familiar, and it would be better if it were possible to give simply a reference. However, the closest reference we know about is Section 2 in Khmaladze and Koul (2004), and, may be, now Section 4 in Khmaladze (2020), yet both will require some work from a reader to match the situations in these papers and in here. It, therefore, is more convenient for readers to see the main steps given explicitly here.

Consider a d-dimensional vector function of the partial derivatives

$$\dot{m}_\theta(x) = (\dot{m}_{\theta k}(x))_{k=1}^d,$$

where

$$\dot{m}_{\theta k}(x) = \frac{\partial}{\partial \theta_k} m_\theta(x),\ k = 1, \ldots, d.$$

Then $(\dot{m}_\theta(X_i))_{i=1}^n$ is $d \times n$-matrix, with d rows and n columns. We assume that for every θ coordinates of $\dot{m}_\theta(x)$ are linearly independent as functions of x, which heuristically means that the model does not include unnecessary parameters.

Obvious example when this condition is true is given by polynomial regression

$$m_\theta(x) = \theta_1 p_1(x) + \theta_2 p_2(x) + \cdots + \theta_d p_d(x), \tag{12}$$

where $p_j(x), j = 1, \ldots, d$, may form a system of (orthogonal) polynomials, or splines (see, for example, Harrell 2015, Sec.2.4.3), or trigonometric polynomials. Being linear in parameters, this model is not essentially different from the model considered in the previous section. Its other form, also frequently used, is given by

$$\mathsf{m}_\theta(x) = \exp[\theta_1 p_1(x) + \theta_2 p_2(x) + \cdots + \theta_d p_d(x)].$$

There are many other examples where $m_\theta(x)$ satisfies differentiability assumption.

Let $\hat{\theta}$ denote the least square estimator of θ, which is an appropriate solution of the least squares' equation

$$\sum_{i=1}^n \dot{m}_{\hat{\theta}}(X_i)\big[Y_i - m_{\hat{\theta}}(X_i)\big] = 0.$$

Without digressing to exact justification (which can be found, for example, in Bates and Watts 2007) assume that Taylor expansion in θ is valid and that together with normalisation by $\sqrt{n}$ it leads to

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{m}_\theta(X_i)\big[Y_i - m_\theta(X_i)\big] - R_n \sqrt{n}(\hat{\theta} - \theta) + \rho_n = 0$$

with a non-degenerate $d \times d$-matrix R_n,

$$R_n = \frac{1}{n} \sum_{i=1}^n \dot{m}_\theta(X_i) \dot{m}_\theta^T(X_i) = \int \dot{m}_\theta(x) \dot{m}_\theta^T(x) \mathrm{d}F_n(x),$$

where ρ_n denotes a d-dimensional vector, such that $E\|\rho_n\|^2 \to 0, n \to \infty$. From the previous display, we obtain asymptotic representation for $\hat{\theta}$:

$$\sqrt{n}(\hat{\theta}-\theta)=R_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\dot{m}_\theta(X_i)\big[Y_i-m_\theta(X_i)\big]+o_P(1).$$

As the next step, expand the differences $\hat{\epsilon}_i = Y_i - m_{\hat{\theta}}(X_i)$ in θ up to linear term and substitute the expression for $\sqrt{n}(\hat{\theta}-\theta)$ to get

$$\hat{\epsilon}_i=\epsilon_i-\dot{m}_\theta^T(X_i)R_n^{-1}\frac{1}{n}\sum_{j=1}^{n}\dot{m}_\theta(X_j)\epsilon_j+o_P(1).$$

In vector form, this becomes

$$\hat{\epsilon}=\epsilon-\dot{m}_\theta^T R_n^{-1}\frac{1}{n}\langle\dot{m}_\theta,\epsilon\rangle+o_P(1), \tag{13}$$

an expression directly analogous to (6). It also describes the vector of residuals as being asymptotically projection of the vector of errors ϵ, parallel to n-dimensional vectors of derivatives

$$(\dot{m}_{\theta 1}(X_i))_{i=1}^n,\dots,(\dot{m}_{\theta d}(X_i))_{i=1}^n.$$

It will be notationally simpler, while computationally not difficult, to change these linearly independent vectors to orthonormal vectors. Namely, introduce the vector function

$$\mu_\theta(x)=R_n^{-1/2}\dot{m}_\theta(x),$$

and then from each of its coordinate function $\mu_{\theta k}(x)$ form the vector

$$\mu_{\theta k,i}=\frac{1}{\sqrt{n}}\mu_{\theta k}(X_i),\ i=1,\dots,n. \tag{14}$$

The two notations are convenient each in its place: $\mu_{\theta k}$ as a vector in $\mathbb{R}^n$ will be useful in expressions like (15), and $\mu_{\theta k}(\cdot)$ as a function in $L_2(F_n)$ will be useful in integral expressions like (16). Their respective norms are equal:

$$\sum_{i=1}^{n}\mu_{\theta k,i}^2=\int\mu_{\theta k}^2(x)\mathrm{d}F_n(x).$$

Which of these two objects we use will be visible in notation and clear from the context.

Now we can write (13) as

$$\hat{\epsilon}=\epsilon-\sum_{k=1}^{d}\mu_{\theta k}\langle\mu_{\theta k},\epsilon\rangle+o_P(1), \tag{15}$$

where the non-vanishing part on the right-hand side is the projection of ϵ orthogonal to vectors $(\mu_{\theta k})_{k=1}^d$. As a consequence, one can show that the following analogue of the representation (8) is true:

$$\begin{aligned}\hat{w}_n(x) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [Y_i - m_{\hat{\theta}}(X_i)] \mathbb{1}_{(X_i \leq x)} \\ &= w_n(x) - \sum_{k=1}^{d} \int_{z \leq x} \mu_{\theta k}(z) \mathrm{d}F_n(z) \int \mu_{\theta k}(z) w_n(\mathrm{d}z) + o_P(1).\end{aligned} \tag{16}$$

This, again, describes $\hat{w}_n$ as asymptotically projection of w_n orthogonal to the functions $(\mu_{\theta k}(\cdot))_{k=1}^{d}$. We are now ready to describe rotation of this projection to another, standard projection, and of $\hat{\epsilon}$ to a vector of another residuals.

Coming back to the linear (in θ) example (12), one finds that $(p_k(\cdot))_{k=1}^{d}$ is the vector of partial derivatives and the matrix R_n becomes

$$R_n = \left(\int p_j(x) p_k(x) dF_n(x) \right)_{j,k=1,\dots,d}$$

and neither of them depend on θ.

3.2 Unitary transformation of $\hat{w}_n$

With some freedom of speech, we say that one can choose the new residuals in any way one wishes; for example, choose them independent of any covariates. Indeed, it would be immediate to choose some d orthonormal n-vectors $(r_k)_{k=1}^{d}$ and then use projection of ϵ, orthogonal to these vectors as new residuals.

To construct the vectors $(r_k)_{k=1}^{d}$, it would be convenient to start with a system of orthonormal polynomials, say, on [0, 1], which are continuous and bounded functions. It often will be convenient to choose the first function r_1 to be identically equal 1. Then choose r_k as vectors with coordinates

$$r_{ki} = \frac{1}{\sqrt{n}} r_k(\frac{i}{n}),\ i = 1, \dots, n. \tag{17}$$

Note, however, that the orthogonality condition

$$\int_0^1 r_k(s) r_l(s) \mathrm{d}s = \int_0^1 r_k(F(z)) r_l(F(z)) \mathrm{d}F(z) = \delta_{k,l},$$

and boundedness and continuity of our functions imply

$$\int_0^1 r_k(F_n(z)) r_l(F_n(z)) \mathrm{d}F_n(z) \to \delta_{k,l}, \quad n \to \infty.$$

This means that the vectors $(r_k)_{k=1}^{d}$ will not be exactly but only asymptotically orthogonal. Small corrections, asymptotically negligible for $n \to \infty$, will be needed, formally. If we insert these corrections in our notation, it will make the text more complicated without opening any new feature of the transformation we want to discuss. Therefore in notations we will identify orthogonal polynomials in continuous time with those, orthonormal on the grid $\{1/n, 2/n, \dots, 1\}$.

It is, certainly, easy to imagine a unitary operator which maps functions $(\mu_{\theta k}(\cdot))_{k=1}^d$ into functions $(r_k(\cdot))_{k=1}^d$, but in practical regression problems we need to calculate it explicitly. We present the operator K below as a product of one-dimensional unitary operators. This allows coding of K recursively, in a loop, which was tried for the case of contingency tables with about 30-dimensional parameter in Nguyen (2017).

Suppose in one-dimensional unitary operator $U_{a,b}$ of Sect. 1 we choose $a = \mu_{\theta 1}$ and $b = r_1$ and apply the resulting operator $U_{\mu_{\theta 1}, r_1}$ to the vector r_2:

$$U_{\mu_{\theta 1}, r_1} r_2 = \tilde{r}_2.$$

Then, the product

$$K_2 = U_{\mu_{\theta 2}, \tilde{r}_2} \times U_{\mu_{\theta 1}, r_1}$$

is unitary operator which maps vectors r_1, r_2 to vectors $\mu_{\theta 1}, \mu_{\theta 2}$ and vice versa, and leaves vectors orthogonal to these four vectors unchanged. For a general k, define $\tilde{r}_k$ as

$$K_{k-1} r_k = \tilde{r}_k, \quad k = 2, \ldots, d.$$

Lemma 1 *The product*

$$K_d = U_{\mu_{\theta d}, \tilde{r}_d} \times \cdots \times U_{\mu_{\theta 1}, r_1}$$

is the unitary operator which maps $(r_k)_{k=1}^d$ to $(\mu_{\theta k})_{k=1}^d$, *and leaves vectors orthogonal to* $(r_k)_{k=1}^d$ *and* $(\mu_{\theta k})_{k=1}^d$ *unchanged.*

The proof of this lemma was given, for example, in Khmaladze (2016), Section 3.4. It may be of independent interest for statistics of directional data, when explicit expression for rotations is needed. At the end of this section, we give an essentially shorter proof.

Thus, the adjoint operator K_d^T is the inverse of K_d and in proposition below we denote

$$\hat{e} = K_d^T \hat{\epsilon}, \tag{18}$$

and recall that X_i-s are numbered in increasing order. We also say

$$E\hat{\epsilon}\hat{\epsilon}^T \sim I - \sum_{k=1}^d \mu_{\theta k} \mu_{\theta k}^T$$

in the sense that for any sequence of n-vectors b_n, such that $\langle b_n, b_n \rangle \to c < \infty$

$$E\langle b_n, \hat{\epsilon} \rangle^2 \sim \langle b_n, b_n \rangle - \sum_{k=1}^d \langle b_n, \mu_{\theta k} \rangle^2, \; n \to \infty.$$

This notion of equivalence is used in the proposition below.

Proposition 2 *Suppose the regression function $m_\theta(x)$ is regular, in the sense that, for every θ, the matrix R_n is of full rank and converges to a matrix R of full rank, and* (15) *is true. Suppose the orthonormal functions $r_k(\cdot), k = 1, \ldots, d$, are continuous and bounded on* $[0, 1]$.*Then*

(i) *for the covariance matrix of residuals $\hat{e}$ the following is true*:

$$E\hat{e}\hat{e}^T \sim I - \sum_{k=1}^{d} r_k r_k^T,\ n \to \infty;$$

(ii) *for the empirical regression process, based on residuals $\hat{e}$ of* (18),

$$\hat{w}_{n,e}(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{e}_i \mathbb{1}_{(X_i \le x)},$$

the following convergence of the covariance function is true:

$$E\hat{w}_{n,e}(x)\hat{w}_{n,e}(y) \to F(\min(x, y)) - \sum_{k=1}^{d} Q_k(F(x))Q_k(F(y)),\ \text{as}\ \ n \to \infty,$$

where $Q_k(t) = \int_0^t r_k(s)\mathrm{d}s$; moreover,

(iii) *the process $\hat{w}_{n,e}$, with time change $t = F(x)$ converges in distribution to projection of standard Brownian motion on* $[0, 1]$ *orthogonal to functions $r_k(\cdot), k = 1, \ldots, d$.*

Proof To prove (i), we do not need the explicit form of the operator K_d, and instead note that according to (15), up to asymptotically negligible term, $\hat{\epsilon}$ is projection of ϵ, orthogonal to collection of n-vectors $\mu_{\theta 1}, \ldots, \mu_{\theta d}$. According to the lemma above, these vectors are mapped by operator K_d^T to n-vectors $r_1, \ldots, r_d$, and the operator K_d is unitary. Therefore, the vector $\hat{\epsilon}$ will be mapped into the vector which, up to asymptotically negligible term, will behave as projection of ϵ, orthogonal to $r_1, \ldots, r_d$:

$$\hat{e} \overset{d}{=} \epsilon - \sum_{k=1}^{d} r_k \langle r_k, \epsilon \rangle + o_P(1). \tag{19}$$

The covariance matrix of the main part on the right side here is the expression given in (i).

To prove (ii), replace $\hat{e}$ by its main term in (19) in the expected value

$$E\hat{w}_{n,e}(x)\hat{w}_{n,e}(y) = \frac{1}{n} E\langle \mathbb{1}_x, \hat{e} \rangle \langle \hat{e}, \mathbb{1}_y \rangle \sim \frac{1}{n} \mathbb{1}_x^T (I - \sum_{k=1}^{d} r_k r_k^T) \mathbb{1}_y.$$

Here, since every $r_k(\cdot)$ is continuous and bounded,

$$\begin{aligned}\frac{1}{\sqrt{n}}\mathbb{1}_x^T r_k &= \frac{1}{n}\sum_{i=1}^{n} r_k(\frac{i}{n})\mathbb{1}_{(X_i\le x)}\\ &= \int_{z\le x} r_k(F_n(z))\mathrm{d}F_n(z) \sim \int_{z\le x} r_k(F(z))\mathrm{d}F(z).\end{aligned}$$

Statement (iii) of convergence in distribution follows not from unitarity property of K_d as such, but from simplicity of its structure, reflected by (19). We have

$$\hat{w}_{n,e}(x) \sim \frac{1}{\sqrt{n}}\langle \mathbb{1}_x, \epsilon - \sum_{j=1}^{d} r_j\langle r_j, \epsilon\rangle\rangle = \frac{1}{\sqrt{n}}\langle \mathbb{1}_x, \epsilon\rangle - \frac{1}{\sqrt{n}}\sum_{k=1}^{d}\langle \mathbb{1}_x, r_k\rangle\langle r_k, \epsilon\rangle.$$

The first inner product on the right side, denoted $w_n(x)$ in (7), converges in distribution to F-Brownian motion. Expression for $\langle \mathbb{1}_x, r_k\rangle$ we considered above, while

$$\langle r_k, \epsilon\rangle = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} r_k(\frac{i}{n})\epsilon_i = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} r_k(F_n(X_i))\epsilon_i = \int r_k(F_n(x))w_n(\mathrm{d}x).$$

Thus, overall representation of $\hat{w}_{n,e}$ through w_n has the form

$$\hat{w}_{n,e}(x) \sim w_n(x) - \sum_{k=1}^{d}\int_{z\le x} r_k(F_n(z))\mathrm{d}F_n(z)\int r_k(F_n(x))w_n(\mathrm{d}x). \tag{20}$$

Since w_n converges in distribution to the F-Brownian motion w_F, which in time $t = F(x)$ becomes a standard Brownian motion w on [0, 1], we see that the process $\hat{w}_{n,e}$ converges in distribution to the Gaussian process given by the right-hand side of (20), which in time $t = F(x)$ can be written as

$$\hat{w}(t) = w(t) - \sum_{k=1}^{d} Q_k(t)\int r_k(s)w(\mathrm{d}s).$$

This is an orthogonal projection of w orthogonal to the functions $r_j(\cdot), j = 1, \dots, d$. □

Proof of Lemma 1 Suppose $K_{l-1}r_j = \mu_{\theta j}$, $1 \le j \le l-1$; then, it follows that $\tilde{r}_l \perp \mu_{\theta j}$, because $r_l \perp r_j$, and operator K_{l-1} is unitary. But then , by its construction,

$$K_l r_j = U_{\mu_{\theta l},\tilde{r}_l} K_{l-1} r_j = U_{\mu_{\theta l},\tilde{r}_l}\mu_{\theta j} = \mu_{\theta j},$$

while

$$K_l r_l = U_{\mu_{\theta l},\tilde{r}_l}\tilde{r}_l = \mu_{\theta l}.$$

Then the rest follows by induction. □

To sum up the result of this section, consider testing the model

$$Z = \gamma_1 r_1 + \cdots + \gamma_d r_d + \alpha. \tag{21}$$

Here errors α have the same distribution as errors ϵ in the original problem, while $(r_k)_{k=1}^d$ are the same fixed n-vectors as the ones we considered above, and $(\gamma_k)_{k=1}^d$ are unknown coefficients. There are no covariates, and therefore, it is not exactly a regression problem. However, any one of such problems can be used as a fixed "standard" problem for testing regression with regular regression functions m_θ. Indeed, suppose one uses least square estimators $(\hat{\gamma}_k)_{k=1}^d$ and then produces the residuals

$$\hat{\alpha} = Z - \hat{\gamma}_1 r_1 - \cdots - \hat{\gamma}_d r_d.$$

Direct calculation shows that they have the same covariance matrix as the one given in (i) of Proposition 2. In other words, we construct residuals $\hat{\epsilon}$ in the original problem of interest, transform them into $\hat{e}$ and then use $\hat{e}$ as if we have $\hat{\alpha}$.

4 The case of multidimensional covariates

It is an important case when the covariate is a finite-dimensional vector. Let us use p for dimension of each X_i. Again, we will not assume anything about probabilistic nature of these covariates, except that

$$F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{\{X_i \le x\}} \to F(x), \quad x \in \mathbb{R}^p,$$

where F is an absolutely continuous distribution function in $\mathbb{R}^p$.

For p-dimensional time, we could have shown that (16) in the previous section is still correct: the regression empirical process $\hat{w}_n$ asymptotically still is the projection of Brownian motion, with the functions $\mu_{\theta k}(\cdot)$-s defined in the same way.

For distribution-freeness of the vector of new residuals $\hat{e}$ it does not matter how do we realise the vectors $(r_k)_{k=1}^d$. They only should stay unchanged no matter what is the regression function m_θ and, hence, the functions $\mu_{\theta k}$-s, provided only that the regularity conditions used in Sect. 3.1 are still valid. They also should not depend on the values of the "physical" covariates X_i-s. Given this, we can come to the construction of the process $\hat{w}_{n,e}$ with only minimal change. For example, one can construct $(r_k)_{k=1}^d$ in literary the same way as earlier, through the functions on [0, 1]. However, in p-dimensional case it will be very natural to use continuous and bounded orthogonal functions $r_k(\cdot)$ of p variables instead. Therefore, we need to find a natural way, similar to (17), to connect functions $r_k(\cdot)$ and vectors r_k.

Even after this is done and we obtain distribution-free residuals, the limit in distribution of the process $\hat{w}_{n,e}$, cf. (ii) of Proposition 2, will not be distribution-free; it will be the corresponding projection of F-Brownian motion w, while now we do not have convenience of the time transformation $t = F(x)$. One can apply to $\hat{w}_{n,e}$ the scanning martingale approach of Khmaladze (1993) or Khmaladze and Koul (2004) or use unitary transformation suggested in Khmaladze (2016) to map the projection (16) into another "standard" projection, in these cases one will need to use estimator of the density of F.

However, it would be preferable to continue viewing covariates $(X_i)_{i=1}^n$ just as given vectors, with not known F and with no probabilistic assumptions, like their i.i.d.-ness. Thus, we suggest to use the approach offered by the theory of optimal transport.

To do this, let us generate an i.i.d. sequence $(\xi_i)_{i=1}^n$ of random variables uniformly distributed on $[0,1]^p$. One could speak here about some distribution G instead of the uniform distribution, but it will be a trite generality. The random variables $(\xi_i)_{i=1}^n$ will not be used to randomise our procedure but to serve as an "anchor", which covariates $(X_i)_{i=1}^n$ will be connected to.

More specifically, consider a one-to-one map T of $(X_i)_{i=1}^n$ to $(\xi_i)_{i=1}^n$, so that $T(X_i) = \xi_j$ for one and only one j, cf. Peyré and Cuturi (2019), Sec. 2.2. There are $n!$ choices of T. Out of them choose the map T_0, which minimises the following sum

$$\sum_{i=1}^n \|X_i - T(X_i)\|,$$

which is the "total distance travelled" of all X_i to $T(X_i)$. Suppose now the n-vectors $(r_k)_{k=1}^d$ are formed as

$$r_{k,i} = \frac{1}{\sqrt{n}} r_k(T_0(X_i)), \quad i = 1, \dots, n. \tag{22}$$

Here $(r_k(\cdot))_{k=1}^d$ is a system of orthonormal functions on $L_2[0,1]^p$. With this choice of $(r_k)_{k=1}^d$, define residuals $\hat{e}$ again as (18).

For any map T,

$$G_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{(T(X_i)\le x)} = \frac{1}{n}\sum_{i=1}^n \mathbb{I}_{(\xi_i \le x)}, \tag{23}$$

so that G_n will converge to the uniform distribution function on $[0,1]^p$.

Using T_0, we can transform the process $\hat{w}_{n,e}$ of Proposition 2 (*ii*) as follows:

$$T_0^* \hat{w}_{n,e}(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \hat{e}_i \mathbb{I}_{(T_0(X_i)\le x)}, \tag{24}$$

where the construction of $\hat{e}$ incorporates, as we said, $T_0(X_i)$-s. The following comment is intended as justification of the use of T_0. Namely, it is not necessary to use minimiser T_0 to produce the version of regression empirical process with standard covariance operator—any T will achieve this. However, if the null hypothesis (11) on the form of the regression function is not correct, expected values of residuals $\hat{e}$ are not zero, but will be, for each contiguous converging alternatives, close to some function, say, h, specific to the alternative (see, for example, Khmaladze and Koul 2004, Sec. 1, or Hajek and Sidak 1967). It will be desirable that the shift of transformed process $T_0^* \hat{w}_{n,e}$ preserves the main pattern present in the shift function h. For this, it is necessary that the transformation of $\hat{w}_{n,e}$ be smooth. One can say that the T should minimise the sum

$$\sum_{i=1}^{n} |h(X_i) - h(T(X_i))|.$$

However, very wide class of alternatives, and therefore, of functions h is apriori possible. The choice of T should not be hinged on a particular h but should be as "smooth" map of $(X_i)_{i=1}^n$ into $(\xi_i)_{i=1}^n$ as possible.

We formulate the next proposition for readers' convenience. It does not require a new proof, and we will give only short comments at the end of it.

Proposition 3 *Suppose the regression function $m_\theta(x)$ is regular, in the same sense as in* Proposition 2. *Suppose the orthonormal functions $r_k(\cdot), k = 1, \dots, d$, are continuous and bounded on* $[0,1]^p$ *and residuals $\hat{e}$ are constructed using the vectors $r_k, k = 1, \dots, d$, defined in* (22)*. Then, as $n \to \infty$,*

(i) *for the covariance matrix of the residuals $\hat{e}$ of* (18) *the following is true*:

$$E\hat{e}\hat{e}^T \sim I - \sum_{j=1}^{d} r_k r_k^T;$$

(ii) *for the empirical regression process, based on residuals $\hat{e}$,*

$$T_0^* \hat{w}_{n,e}(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{e}_i \mathbb{1}_{(T_0(X_i) \le x)},$$

the following convergence of the covariance function is true: as $n \to \infty$,

$$ET_0^* \hat{w}_{n,e}(x) T_0^* \hat{w}_{n,e}(y) \to G(\min(x,y)) - \sum_{k=1}^{d} Q_k(x) Q_k(y),$$

where $Q_k(x) = \int_{z \le x} r_k(z)\mathrm{d}z$;
moreover,

(iii) *the process $T_0^* \hat{w}_{n,e}$ converges in distribution to projection of standard Brownian motion on* $[0,1]^p$ *orthogonal to functions $r_k(\cdot), k = 1, \dots, d$.*

Given two orthonormal systems of n-vectors $(\mu_{\theta k})_{k=1}^d$ and $(r_k)_{k=1}^d$, the operator K_d will rotate one system into another, regardless of how these systems have been constructed. Therefore, (19) is also true for p-dimensional time, and this implies (i).

To see that (iii) is true denote $\mathbb{1}_{T_0,x}$ the vector with coordinates $\mathbb{1}_{(T_0(X_i) \le x)}$. Now we use (19) to write the process $T_0^* \hat{w}_{n,e}$ in the form

$$T_0^* \hat{w}_{n,e}(x) \sim \frac{1}{\sqrt{n}} \langle \mathbb{1}_{T_0,x}, \epsilon - \sum_{k=1}^{d} r_k \langle r_k, \epsilon \rangle \rangle,$$

and then use the representation of $(r_k)_{k=1}^d$ through functions $(r_k(\cdot))_{k=1}^d$:

$$\frac{1}{\sqrt{n}}\langle \mathbb{1}_{T_0,x}, \epsilon\rangle = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbb{1}_{(T_0(X_i)\le x)}\epsilon_i = T_0^* w_n(x)$$

$$\frac{1}{\sqrt{n}}\langle \mathbb{1}_{T_0,x}, r_k\rangle = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(T_0(X_i)\le x)} r_k(T_0(X_i)) = \int r(z)dG_n(z),$$

so that

$$\frac{1}{\sqrt{n}}\langle \mathbb{1}_{T_0,x}, r_k\rangle\langle r_k, \epsilon\rangle = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(T_0(X_i)\le x)} r_k(T_0(X_i))\frac{1}{\sqrt{n}}\sum_{i=1}^{n} r(T_0(X_i))\epsilon_i$$
$$= \int_0^x r(z)dG_n(z)\int r_k(z)T_0^* w_n(\mathrm{d}z).$$

This altogether leads to

$$T_0^* \hat{w}_{n,e}(x) \sim T_0^* w_n(x) - \sum_{k=1}^{d}\int_0^x r_k(z)dG_n(z)\int r_k(z)T_0^* w_n(\mathrm{d}z).$$

The process $T_0^* w_n$ obviously converges to G-Brownian motion (that is, standard Brownian motion) on $[0, 1]^p$, while $T_0^* \hat{w}_{n,e}$ differs from it by the term which involves only finitely many linear functionals from it.

We formulated (ii) for the sake of some symmetry of presentation. To see that (ii) is true, one can follow the proof of (ii) in Proposition 2 using (23) in place of $\mathbb{1}_x^T \mathbb{1}_y$,

$$G_n(\min(x, y)) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(T_0(X_i)\le x)}\mathbb{1}_{(T_0(X_i)\le y)} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(T_0(X_i)\le \min(x,y))},$$

and using $\int_{z\le x} r_k(z)dG_n(z)$ in place of $\frac{1}{\sqrt{n}}\mathbb{1}_x^T r_k$. On the other hand, it also follows from (iii).

For optimal transport method to work, one does not need that the points X_i and ξ_j are in the same set. One only needs $n \times n$ matrix of "distances" $\|X_i - \xi_j\|$, or "costs" of transporting X_i to ξ_j. It is also not necessary that $(\xi_i)_{i=1}^n$ be generated as random variables—they can be strategically placed to form a uniformly spread net. On the other hand, to find a minimiser T_0 can be computationally costly, more so than the estimation of density based on F_n. More detailed comparison of the two methods is the subject of the paper Bancolita (2019). In the next section, we choose $p = 2$ and consider several configurations of covariates $(X_i)_{i=1}^n$ in $[0, 1]^2$. We also use an approximation of T_0 by computationally simpler transformation $\tilde{T}$ called Hungarian method, see, for example, Kuhn (1956), Tomizawa (1971).

5 On power considerations

We do not advocate in this paper any particular test. Any test based on a functional from the transformed empirical process $T_0^* \hat{w}_{n,e}(x)$ is asymptotically distribution-free, and which particular functional will be chosen as test statistics remains in discretion of a user.

Figures 2 and 3 provide some illustration regarding distribution-free property of transformed process and statistics based on it. For this, we needed to choose statistics of some classical test; for example, two one-sided Kolmogorov–Smirnov statistics

$$D_n^+ = \max_x \hat{w}_n(x) \quad \text{and} \quad D_{n,e}^+ = \max_x T_0^* \hat{w}_{n,e}(x).$$

One-sided statistics was quicker to calculate and otherwise the choice was immaterial. Figure 2 shows three cases with different distributions of covariates $(X_i)_{i=1}^n$, which produces scatterplots of different patterns, although the sample size was not

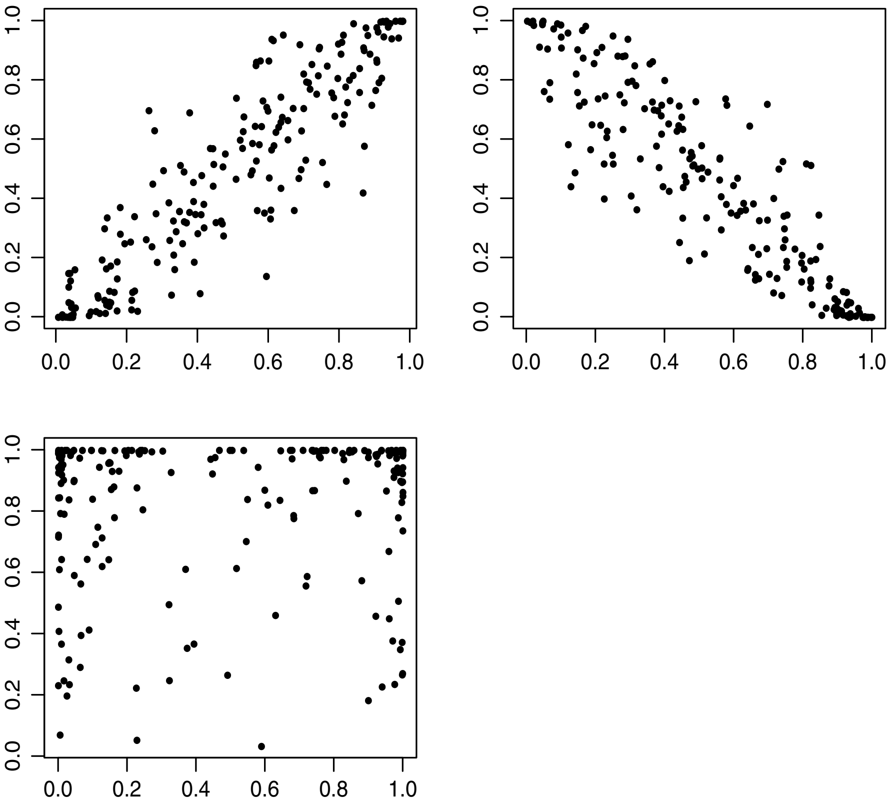

Fig. 2 In the three scatterplots, the covariates $(X_i)_{i=1}^n$ are generated as 2-dimensional iid random vectors, but in the first row coordinates of each X_i are not independent: they are $X_{i1} \sim \mathcal{U}[0,1], X_{i2} \sim Beta(8(1-X_{i1}), 8X_{i1})$ on the left scatterplot, and $X_{i1} \sim \mathcal{U}[0,1], X_{i2} \sim Beta(8X_{i1}, 8(1-X_{i1}))$ on the right one. On the third scatterplot the coordinates are independent, but have different *Beta*-distributions: $X_{i1} \sim Beta(0.35, 0.35)$ and $X_{i2} \sim Beta(0.2, 0.2)$

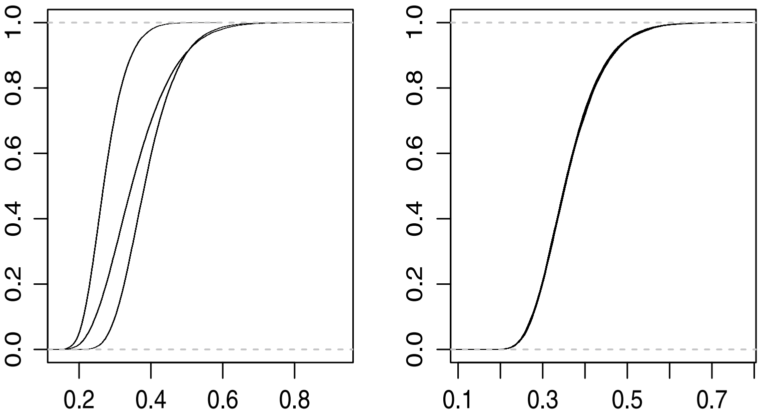

Fig. 3 On the left panel, we show three simulated distribution functions of statistic $D_n^+ = \max_x \hat{w}_n(x)$ for X_i-s distributed as on the three scatterplots shown above. These distribution functions are indeed different. On the right panel, there are also three graphs of distribution functions of the statistic from the transformed process $D_{n,e}^+ = \max_x T_0^* \hat{w}_{n,e}(x)$ for the same three scatterplots. Sample size in all cases was $n = 200$. Visually the graphs are indistinguishable.

too big, $n = 200$. Figure 3 shows that the distribution of statistic D_n in these three cases was quite different, while the distribution of statistic $D_{n,e}$, with covariates of the same three different distributions, is almost undistinguishable.

On the other hand, distribution-freeness can not be the only requirement on a statistic or an underlying empirical process, because trivial and useless choices are possible. The version of regression empirical process constructed in this paper satisfies two requirements, not one: a) under the null hypothesis its limit distribution does not depend on parametric family of regression functions or the true value of the parameter, and b) for any sequence of alternative regression functions b_n, converging to m_θ at some θ from the (functional) direction ϕ,

$$b_n(x) = m_\theta(x) + \frac{1}{\sqrt{n}}\phi_n(x), \qquad \int [\phi_n(x) - \phi(x)]^2 \mathrm{d}F(x) \to 0,$$

the statistic of locally most powerful test for testing against the sequence b_n is a functional of the transformed regression empirical process. So, it is asymptotically distribution-free and sensitive to all local alternatives at the same time.

Indeed, the last claim is true because the regression empirical process $\hat{w}_n$ does have the property b), see, for example, Khmaladze and Koul (2004), and because the process $T_0^* \hat{w}_{n,e}$ is its "smooth" one-to-one transformation. The latter also implies that test statistics based on $\hat{w}_n$ can be viewed as statistics based on $T_0^* \hat{w}_{n,e}$,

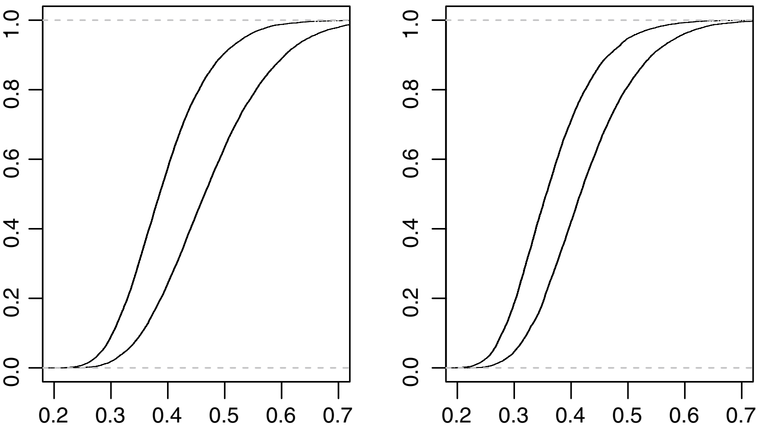

Fig. 4 Here $\phi(x) = x_2^3$ and sample size $n = 200$. Although the uniform distance, and therefore the distance in total variation, between the two distributions on both panels are very similar, the overall impression well may be that D_n^+, the KS statistic from unmodified regression process (left panel) reacts on the alternative somewhat better than $D_{n,e}^+$.

and vice versa. Therefore, at the first glance natural question on power behaviour of the "same test" from the two processes is actually a question of comparing two *different* tests from the same empirical process. This is the case, for example, for two Kolmogorov–Smirnov statistics above, with the second maximum taken from $\hat{w}_{n,e}(x)$ if covariates are one-dimensional. For a reader with some experience in goodness-of-fit theory, it will be clear that both tests are admissible—neither dominates the other in statistical power.

Here is an illustration of this point in two more figures. The left panel in Fig. 4 shows distribution functions of statistic D_n^+ under the null model (the one more to the left), with two-dimensional covariates and with

$$m_\theta(X_i) = \theta_0 + \theta_{10}(X_{1i} - \bar{X}_{1n}) + \theta_{01}(X_{2i} - \bar{X}_{2n}) + \theta_{11}(X_{1i}X_{2i} - \overline{X_1X_2}),$$

and under alternative $m_\theta(x) + x_2^3$, while the right panel shows the distributions of statistic $D_{n,e}^+$ in the same situation. Figure 5 shows what happens under the same model, but now with alternative $m_\theta(x) + \sin(\pi x_2/2)$.

To complement short discussion in the previous section on why we need to use the optimal transport T_0 note the following: as we remarked, the choice of the optimal transport map will transform the shape of the bias term ϕ in a consistent way, but one needs to be sure that this consistency is preserved as $n \to \infty$. This latter is true, however, as it follows, for example, from Cuesta-Albertos et al. (1997) Theorem 3.2.

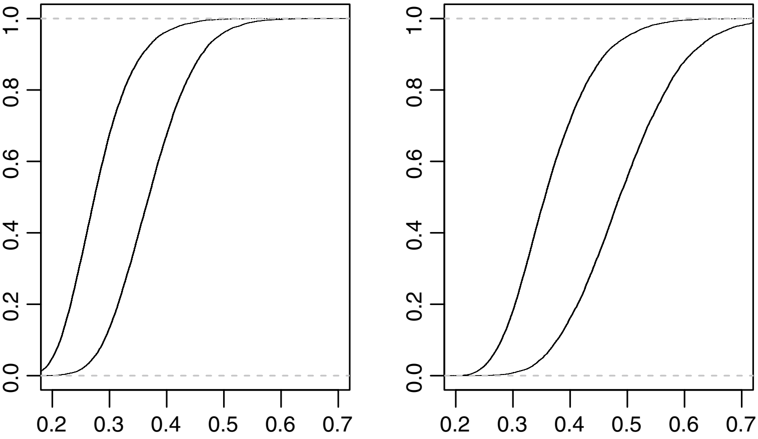

Fig. 5 Here $\phi(x) = \sin(\pi x_2/2)$ and sample size $n = 200$. Although the uniform distance between the two distributions on both panels is still similar, the overall impression is that $D_{n,e}$, the KS statistic from the transformed regression empirical process (right panel) reacts on the alternative better than D_n.

Acknowledgements For the results, shown in Figs. 2, 3, 4 and 5 and many more experiments, not included here, the author is grateful to his student at the time, Richard White. Author is also grateful to the referees for patience with many imperfections of the initial draft and for the number of useful advice. Careful reading by Sara Algeri greatly helped to improve the text at the final stages.

References

Bancolita J. (2019). Numerical investigation of Khmaladze projection approach in regression, Research Report, School of Mathematics and Statistics, Victoria University of Wellington.

Bates, D. M., Watts, D. G. (2007). *Nonlinear regression analysis and its applications*, 2nd ed. New York: Wiley.

Can, S. U., Einmahl, J. H. J., Laeven, R. (2020). Distribution free two-sample test for tail copulas, Work in progress.

Chown, J., Müller, U. U. (2018). Detecting heteroscedasticity in non-parametric regression using weighted empirical processes. *Journal Royal Statistical Society, B, 80,* 961–974.

Cook, R., Weisberg, S. (1982). *Residuals and influence in regression*. Boca Raton: Chapman and Hall.

Cuesta-Albertos, J. A., Matrán, C., Tuego-Diaz, A. (1997). Optimal transportation plans and convergence in distribution. *Journal Multivariate Analysis, 60,* 72–83.

de Valk, C., Segers, J. (2018). Stability and tail limits of transport-based quantile contours, arXiv:1811.12061.

del Barrio, E., Cuesta-Albertos, J. A., Hallin, M., Matrán, C. (2018). Center-outward distribution functions, quantiles, ranks, and signs in R^d, arXiv:1806.01238.

Dette, H., Hetzler, B. (2009). A simple test for the nonparametric form of the variance function in nonparametric regression. *Annals Institute Statistical Mathematics, 61,* 861–886.

Dette, H., Munk, A. (1998). Testing heteroscedasticity in nonparametric regression. *Journal Royal Statistical Society, B, 60,* 693–708.

Dette, H., Neumeyer, N., Van Keilegom, I. (2007). A new test for the parametric form of the variance function in non-parametric regression. *Journal Royal Statistical Society, B, 69,* 903–917.

Einmahl, J. H. J., Khmaladze, E. V. (2001). Two-sample problem in R^m and measure-valued martingales, State of the Art in Statistics and Probability Theory; Festschrift for Willem R. van Zwet. *IMS Lecture Notes-Monograph Series, 36,* 434–464.

Gonzalez-Manteiga, W., Crujeiras, R. M. (2013). An updated review of Goodness-of-Fit tests for regression models. *TEST, 22,* 361–447.

Hajek, J., Sidak, Z. (1967). *Theory of rank tests*. New York: Academic Press.

Harrell, F. E., Jr. (2015). *Regression modelling strategies*, 2nd ed. Berlin: Springer.

Khmaladze, E. V. (1993). Goodness of fit problems and scanning innovation martingales. *Annals of Statistics, 21,* 798–829.

Khmaladze, E. V. (2013). Note on distribution free testing for discrete distributions. *Annals of Statistics, 41,* 2979–2993.

Khmaladze, E. (2016). Unitary transformations, empirical processes and distribution free testing. *Bernoulli, 22,* 563–588.

Khmaladze, E. V. (2020). Projection approach to distribution-free testing for point processes. Regular models. *Transactions of A. Razmadze Mathematical Institute, 174,* 155–176.

Khmaladze, E. V., Koul, H. L. (2004). Martingale transforms goodness of fit tests in regression models. *Annals of Statistics, 32,* 955–1034.

Koenker, R. (2005). *Quantile regression*. New York: Cambridge University Press.

Koul, H. L., Müller, U. U., Schick, A. (2017). Estimating the error distribution in a single-index model, *From statistics to mathematical finance, festschrift in honour of Winfried stute,* In Dietmar Ferger, Wenceslao Gonzalez-Manteiga, Thorsten Schmidt, Jane-Ling Wang (Eds.) Springer, Heidelberg, Berlin.

Kuhn, H. W. (1956). Variants of the Hungarian method for assignment problems. *Naval Research Logistics Quarterly, 3,* 253–258.

McCullagh, P., Nelder, J. A. (2008). *An introduction to generalized linear models*, 3rd ed. Boca Raton: Chapman & Hall/ CRC Monographs on Statistics.

Müller, U. U., Schick, A., Wefelmeyer, W. (2009). Estimating the innovation distribution in nonparametric autoregression. *Probability Theory Related Fields, 144,* 53–77.

Nguyen, T. T. M. (2017). New approach to distribution free tests in contingency tables. *Metrika, 80,* 153–170.

Peyré, G., Cuturi, M. (2019). Computational optimal transport, arXiv:1803.00567v2 [statML].

Roberts, L. (2019). On distribution free goodness of fit testing of Bernoulli trials. *Statistics and Probability Letters, 150,* 47–53.

Stute, W. (1997). Nonparametric model checks for regression. *Annals of Statistics, 25,* 613–641.

Tomizawa, N. (1971). On some techniques useful for solution of transportation network problems. *Networks, 1*(2), 173–194.

Villani, C. (2009). *Optimal transport: Old and new*. Berlin: Springer.