Distribution free testing for linear regression. Extension to general parametric regression

E. Khmaladze

Abstract: Recently a distribution free approach for testing parametric hypotheses based on unitary transformations has been suggested in [Khmaladze(2013), Khmaladze(2016), Khmaladze(2017)] and further studied in [Roberts(2019)] and [Nguyen(2017)]. In this paper we show, how can the approach be applied to distribution free testing of linear regression. Namely, the empirical processes suggested in this paper have two properties: their asymptotic null distribution depends neither on covariates, nor on the distribution of i.i.d. errors, and they are sensitive to all contiguous alternatives to the hypothetical family of regression functions.

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 $(X_j)_{j=1}^n$, the errors $(\epsilon_i)_{i=1}^n$ are independent, and have expected value zero and finite variances – for the sake of simplicity of presentation we assume these variances all 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.

Our aim is to describe a new method to build asymptotically distribution free theory for testing such hypothesis. 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 is based on the application of unitary operators as described more or less recently in [Khmaladze(2013), Khmaladze(2016), Khmaladze(2017)] and studied in [Roberts(2019)] and [Nguyen(2017)].

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 – a mild 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,

$$\hat\epsilon = Y - X\hat\theta \quad \text{with} \quad \hat\theta = \langle Y, z\rangle,$$

where $z = X/\langle X, X\rangle^{1/2}$ is the normalised vector of covariates. The natural object to base a goodness of fit test upon is given by the partial sums process (see, e.g., [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 not the Brownian bridge. Its distribution remains dependent on behaviour of the covariates. In particular, the limit distribution of 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 of unit norm: b with $\|a\| = \|b\| = 1$. If $a = b$ we take $U_{a,b} = I$. This operator in unitary, it maps a into b and b into a, and it maps any vector c, orthogonal to a and b, to itself, see, e.g., [Khmaladze(2013)], Sec. 2. Now choose $a = z$ and choose b equal $r = (1,\dots,1)^T/\sqrt{n}$, the vector not depending on covariates at all. Since the vector of residuals $\hat{\epsilon}$ is orthogonal to the vector z, we obtain:

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

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

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

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

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

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

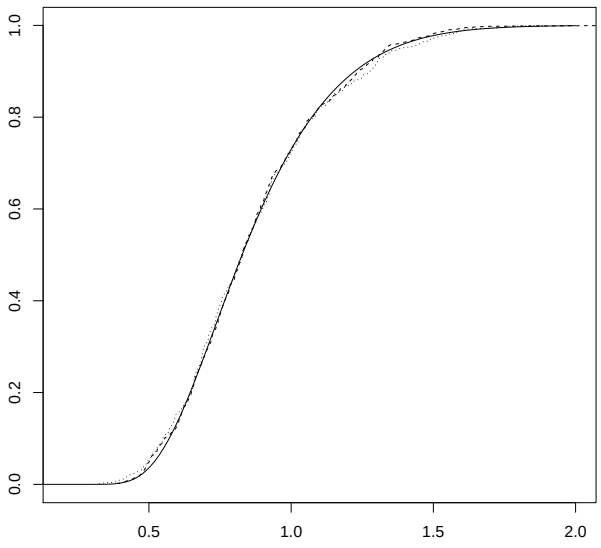

Figure 1: The smooth line is Kolmogorov distribution function. The two other ones are simulated distributions of $\max_x |\hat{w}_{n,e}(x)|$ for two entirely different behaviour of covariates. In one case X_i-s have uniform distribution on $[0, 2]$ while in the other they have Gaussian distribution $N(1,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

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

forms of residuals have been suggested, various decompositions of z, especially when covariates X_i are multidimensional, have been studied and intricate approximations, mostly for quadratic forms from $\hat{\epsilon}$, have been developed. Here assumption of normality, arbitrary as it could have been in many cases, has been made more or less casually. If one is allowed somewhat free speech, one could say that a delicate mathematical lace have been created. Very good source for this material is the book [Cook, Weisberg (1982)]. As to well known modern guide in the existing theory we refer to [McCullagh, Nelder(2008)]. The most recent review on goodness of fit problems in regression, which we know of, is [Gonzales Manteiga, Crujeiras (2013)].

In conclusion of this section we make one more remark of general character. The partial sum processes, like $\hat{w}_n$ represent one of the main objects of asymptotic theory. However, it is often that a somewhat different form of such processes is considered, one simple example of which would be

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

(see more sophisticated form of the weight function in recent paper [Chown, 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, Munk (1998)], [Dette, Hetzler (2009)], [Dette *et al* (2007)] and , and loc.cit. [Chown, Müller (2018)] is the problem of testing heterogeneity of errors. The same scanning is basically unavoidable in study 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 $1, 2, \ldots, n$, it would be strange not to examine the sequence $\hat{\epsilon}_1, \hat{\epsilon}_2, \ldots, \hat{\epsilon}_n$, 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 $\hat{w}_{ne}$.

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 the next Section 2 we consider the general form of one-dimensional linear regression. In the following Section 3 we consider general parametric regression. In this case the time transformation, considered in (iii) of the Proposition 2 below again leads to distribution free-ness if F is continuous. If F is discrete, then the method suggested in [Khmaladze(2013)], section 2, can be easily used. In Section 4 we consider multidimensional X_is. Here one can use the approach of [Khmaladze(2016)], section 2, but it would seems unavoidable to employ density estimation of the covariates. However, will show that a different approach can be used and density estimation can be avoided. This approach we borrowed from the theory of optimal transportation, or Monge - Kantorovich transportation problem, see, e.g., [Villani(2009)]. Very interesting probabilistic applications of this theory have been recently given in [del Bario *et al* (2018)] and [Segers(2018)].

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

Consider the general form of one-dimensional linear regression model

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

which is standard linear regression on real line. The $\mathbf{1}$ here denotes a vector with all coordinates equal to the number 1. Instead of (4) consider its slightly modified but more convenient form

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

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 and notation

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

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

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

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. One can show that the limit distribution of the regression process with these residuals,

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

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

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

Here one can think of h as a continuous time "trace" of $\tilde{z}$.

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

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

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

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

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

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

The third term produces the following expression

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \sum_{j=1}^n \epsilon_j (X_j - \bar{X}) \frac{1}{\sum_{j=1}^n (X_j - \bar{X})^2} \sum_{i=1}^n (X_i - \bar{X}) \mathbb{I}_{(X_i \le x)} \\
&= \int (y - \bar{X}) dw_n(y) \frac{1}{\int (y - \bar{X})^2 dF_n(y)} \int^x (y - \bar{X}) dF_n(y) \\
&= \int h_n(y) dw_n(y) \int^x h_n(y) dF_n(y),
\end{aligned}$$

where

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

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

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

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 $\int y^2 dF(y) < \infty$ (and in our assumption this integral equals 1), then $\hat{w}_\epsilon$ is 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 this will require re-shuffling of our initial pairs of observations, probability measure we work under will not change, because 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

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

where the second equality 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 we do not need any further specification of $\tilde{r}$, but 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

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

and

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

and 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

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

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 $\hat{e}$,*

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

has the covariance function

$$E\hat{w}_{n,e}(x)\hat{w}_{n,e}(y) = F_n(\min(x,y)) - F_n(x)F_n(y) - Q_n(F_n(x))Q_n(F_n(y)) + O(1/n),$$

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

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

(iii) As a corollary of (ii), 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 $U_{\tilde{z},\tilde{r}}$. Therefore

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

Calculation of the covariance matrix of the right hand side is now not difficult using shorthand formulas $E\epsilon\langle \epsilon, a \rangle = a$ and $E\langle \epsilon, a \rangle \langle \epsilon, b \rangle = \langle a, b \rangle$. After some algebra we obtain the expression given in (i).

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

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

Opening the brackets in the last expression one can find that

$$\frac{1}{n}\langle \mathbb{I}_x, \mathbb{I}_y \rangle = F_n(\min(x, y)) \quad \text{and} \quad \frac{1}{n}\langle \mathbb{I}_x, r \rangle \langle \mathbb{I}_y, r \rangle = F_n(x)F_n(y),$$

while

$$\begin{aligned} \frac{1}{n}\langle \mathbb{I}_x, \tilde{r} \rangle \langle \mathbb{I}_y, \tilde{r} \rangle &= \frac{1}{n}\sum_{i=1}^{n} \tilde{r}(\frac{i}{n})\mathbb{I}_{(X_i \le x)} \frac{1}{n}\sum_{i=1}^{n} \tilde{r}(\frac{i}{n})\mathbb{I}_{(X_i \le y)} \\ &= \frac{1}{n}\sum_{i=1}^{nF_n(x)} \tilde{r}(\frac{i}{n}) \; \frac{1}{n}\sum_{i=1}^{nF_n(y)} \tilde{r}(\frac{i}{n}) = Q_n(F_n(x))Q_n(F_n(y)) \end{aligned}$$

which proves (ii).

The statement (iii) follows if we note that the limit of 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 $w(t) - tw(1) - Q(t)\int_0^1 \tilde{r}(s)dw(s)$, which exactly is the projection described in (iii). **Λ**

In both regression models (1) and (5) the process $\hat{w}_n$ turns out to be a projection of a Brownian motion, but for different values of covariates these projections are different. However, it is geometrically clear that it should be possible to rotate one projection into another, and this anther 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 usefulness of this approach depends on how practically simple the rotation will be. For us, the transformations of $\hat{\epsilon}$ into $\hat{e}$ looks very simple.

Finally, note that the model (5) includes two estimated parameters while the 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. Situation when one needs to rotate several vectors at once, as well as general form of parametric regression will be considered in the next Section 3.

3 General parametric regression

Now consider testing regression model

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

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

$$m_\theta(x) = \theta_1 p_1(x) + \theta_2 p_2(x) + \dots + \theta_d p_d(x)$$

where $p_j(x), j = 1, \ldots, d$, may form a system of (orthogonal) polynomials, or splines (see, e.g., [Harrell (2015)], Sec.2.4.3), or trigonometric polynomials. There certainly are also many examples, where $m_\theta(x)$ is not linear in θ.

Now denote

$$\dot{m}_\theta(x) = (\frac{\partial}{\partial\theta_1} m_\theta(x), \ldots, \frac{\partial}{\partial\theta_d} m_\theta(x))^T$$

a d-dimensional vector-function of the partial derivatives. 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.

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

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

Without digressing to exact justification (which can be found, e.g., in [Bates, Watts (2007)]) assume that Taylor expansion in θ is valid and that together with normalization by $\sqrt{n}$ it leads to

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

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

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

and d-dimensional vector of residuals ρ_n, such that $E\|\rho_n\|^2 \to 0, n \to \infty$. Below for the terms asymptotically negligible in the same sense we will use notation $o_P(1)$. From the previous display we obtain asymptotic representation for $\hat{\theta}$:

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

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

$$Y_i - m_{\hat{\theta}}(X_i) = Y_i - m_\theta(X_i) - \dot{m}_\theta^T(X_i) R_n^{-1} \frac{1}{n} \sum_{j=1}^{n} \dot{m}_\theta(X_j)[Y_j - m_\theta(X_j)] + o_P(1)$$

or

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

In vector form this becomes

$$\hat{\epsilon} = \epsilon - \dot{m}_\theta^T R_n^{-1} \frac{1}{n} \langle \dot{m}_\theta, \epsilon \rangle + o_P(1), \tag{12}$$

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 the system of d n-dimensional vectors of derivatives

$$(\frac{\partial}{\partial \theta_1} m_\theta(X_i))_{i=1}^n, \ldots, (\frac{\partial}{\partial \theta_d} m_\theta(X_i))_{i=1}^n.$$

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

$$\mu_{\theta k}(x) = R_n^{-1/2} \frac{\partial}{\partial \theta_k} m_\theta(x), \quad k = 1, \ldots, d,$$

and then the vectors

$$\mu_{\theta k,i} = \frac{1}{\sqrt{n}} \mu_{\theta k}(X_i), \ i = 1, \ldots, n. \tag{13}$$

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 (14), and $\mu_{\theta k}(\cdot)$ as a function in $L_2(F_n)$ will be useful in integral expressions like (15). Their respective norms are equal:

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

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

Now we can write (12) as

$$\hat{\epsilon} = \epsilon - \sum_{k=1}^{d} \mu_{\theta k}^T \langle \mu_{\theta k}, \epsilon \rangle + o_P(1), \tag{14}$$

where the leading term on the right hand side is the projection of ϵ orthogonal to vectors $\mu_{\theta,k}$. As a consequence, one can show that the following analogue of the representation (8) is true:

$$\begin{aligned} \hat{w}_n(x) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [Y_i - m_{\hat{\theta}}(X_i)] \mathbb{I}_{(X_i \le x)} \\ &= w_n(x) - \sum_{k=1}^{d} \int_{z \le x} \mu_{\theta k}(z) dF_n(z) \int \mu_{\theta k}(z) dw_n(z) + o_P(1). \end{aligned} \tag{15}$$

We are ready to describe rotation $\hat{\epsilon}$ to a vector of another residuals.

With some freedom of speech, we say that one can choose these new residuals in any way we wish; for example, choose them independent of any covariates. In particular, let $r_1(\cdot)$ be a function on $[0,1]$, identically equal 1, and with this let vectors r_k be defined as $r_{ki} = r_k(i/n)/\sqrt{n}$, where the system of functions $(r_k(\cdot))_{k=1}^d$ is such that

$$\frac{1}{n} \sum_{i=1}^{n} r_k(\frac{i}{n}) r_l(\frac{i}{n}) = \delta_{k,l}, \ k, l = 1, \dots, d.$$

If we derive a unitary operator K, which maps orthonormal vectors $(\mu_{\theta,k})_{k=1}^d$ into vectors $(r_k)_{k=1}^d$, then this operator will map $\hat{\epsilon}$ into $\hat{e}$, and the covariance matrix of these new residuals will be defined solely by $(r_k)_{k=1}^d$ or $(r_k(\cdot))_{k=1}^d$.

As a side and rather inconsequential remark we note that it would be immediate to choose orthonormal polynomials on $[0,1]$, i.e. such that

$$\int_0^1 r_k(s) r_l(s) ds = \delta_{k,l},$$

which are continuous and bounded functions. Such polynomials will not satisfy the orthogonality condition in the previous display above, but will require small corrections, asymptotically negligible for $n \to \infty$. 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 $\{1/n, 2/n, \ldots, 1\}$.

It is essential that the structure of K allows convenient handling. We present it here as a product of one-dimensional unitary operators. This allows coding of K in a loop, and was tried for the case of contingency tables with about 30-dimensional parameter in [Nguyen(2017)].

Suppose in one-dimensional unitary operator $U_{a,b}$ we choose $a = \mu_{\theta,1}$ and $b = r_1$ and apply the resulting operator $U_{\mu_{\theta,1},r_1}$ to 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 vice versa, and leaves vectors orthogonal to* $(r_k)_{k=1}^d$ *and* $(\mu_{\theta,k})_{k=1}^d$ *unchanged.*

The proof of this lemma is given, e.g., in [Khmaladze(2016)], section 3.4. It may be of independent interest for statistics of directional data, when explicit expression for rotations is needed.

Thus, in proposition below we denote

$$\hat{e} = K_d \hat{\epsilon}, \tag{16}$$

and once again assume 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 $\langle b_n, b_n\rangle \to c < \infty$

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

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 (14) is true. Suppose the 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:

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

(ii) for the empirical regression process, based on residuals $\hat{e}$ of (16),

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

the following convergence of the covariance function is true:

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

where $Q_k(t) = \int_0^t r_(s)ds$;

(iii) moreover, 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 $r_j(\cdot), j = 1, \ldots, d$.

To prove (i) we now avoid using the explicit form of the operator K_d, and instead note that according to (14), 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 to n-vectors $r_1, \ldots, r_d$, and the operator K_d is unitary. It is also continuous. Therefore the vector $\hat{\epsilon}$ will be mapped into the vector which, up to asymptotically negligible term, will be projection of ϵ orthogonal to $r_1, \ldots, r_d$:

$$\hat{e} = \epsilon - \sum_{k=1}^{d} r_k \langle r_k, \epsilon \rangle + o_P(1). \tag{17}$$

And the covariance matrix of this vector is the expression given in (i).

To prove (ii), replace $\hat{e}$ by its main term in (17) in the expected value

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

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

$$\frac{1}{\sqrt{n}} \mathbb{I}_x^T r_k = \frac{1}{n} \sum_{i=1}^{n} r_k(\frac{i}{n}) \mathbb{I}_{(X_i \le x)} \sim \int_{z \le x} r_k(F_n(z)) dF_n(z).$$

Statement (iii) of convergence in distribution follows not from unitarity property of K_d as such, but from simplicity of its structure, reflected by (17). We have

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

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

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

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

$$\hat{w}_{n,e}(x) \sim w_n(x) - \sum_{k-1}^{d} \int_{z \le x} r_k(F_n(z)) dF_n(z) \int r_k(F_n(x)) dw_n(x). \tag{18}$$

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 the display above, which in time $t = F(x)$ can be written as

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

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

4 The case of multi-dimensional 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

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

where F is an absolutely continuous distribution function in $\mathbb{R}^p$. For simplicity of presentation, it will be convenient, however, to assume that F is replaced by its copula function, or, equivalently, F itself is supported on $[0, 1]^p$, although the support can be a proper subset of $[0, 1]^2$.

For p-dimensional time, we could have shown that (15) in the previous section is still correct. One of the relatively familiar ways to obtain distribution-free transformation of this process would be to use the scanning martingale's approach of [Khmaladze and Koul(2004)] to the projection (15). Another possibility would be to

use unitary transformations suggested in [Khmaladze(2016)] to map the projection (15) into another "standard" projection, changing simultaneously the functions $\mu_{\theta k}(\cdot)$ and distribution F to the corresponding objects of our choice. In doing this one will need to use estimator of the density of F, assuming it exists. Here, however, we will see that both tasks can be achieved, again simultaneously but simpler, using the approach suggested by the theory of optimal transport.

For distribution free-ness of the vector of new residuals it does not matter how do we realise the vectors $(r_k)_{k=1}^d$. For example, one can represent them in literary the same way as in (9) – the covariance matrix of the new residuals will depend on $r(\cdot)$ and not on covariates. However, similarly to (13), see also discussion following (15), it will be very natural to connect vectors $(r_k)_{k=1}^d$ with a system of piecewise continuous orthogonal functions $r_k(\cdot)$ of p variables. 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" to connect covariates $(X_i)_{i=1}^n$ to new ones which are uniformly distributed on $[0,1]^p$.

Now consider a one-to-one "push forward" 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, Cuturi (2019)], sec. 2.2. There are $n!$ choices of T. Out of them we choose the map T_0, which minimises the following sum

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

Suppose now the vectors $(r_k)_{k=1}^d$ are formed as

$$r_{k,i} = \frac{1}{\sqrt{n}} r_k(T_0(X_i)), \quad k = 1, \dots, d. \tag{19}$$

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 (16). Justification of the use of the operator T_0 partly

comes from equality

$$G_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{(T_0(X_i)\leq x)} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{(\xi_i \leq x)}, \tag{20}$$

which shows that G_n will converge to the uniform distribution function on $[0,1]^p$. As a corollary of (19) and (20), the behaviour of statistics, which are invariant under permutation, is governed by G_n and not by F_n. For example

$$\frac{1}{n}\sum_{i=1}^{n} r_k(T_0(X_i))\mathbb{I}_{(T_0(X_i)\leq x)} = \int_{z\leq x} r_k(z) dG_n(z). \tag{21}$$

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

$$T_0^* \hat{w}_{n,e}(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \hat{e}_i \mathbb{I}_{(T_0(X_i)\leq x)}, \tag{22}$$

where the construction of $\hat{e}$ incorporates, as we said, $T_0(X_i)$-s. The following comment is intended as further justification of the use of T_0. 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, in the case when the null hypothesis (11) 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, e.g., [Khmaladze and Koul(2004)], sect. 1, or [**?**]). It will be desirable that the shift of transformed process $T_0^* \hat{w}_{n,e}$ preserve 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. Therefore, 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. This motivates the choice of T_0.

In the formulation of the next proposition let us denote $T_0(X_i) = \xi_j$.

Proposition 3. *Suppose the regression function $m_\theta(x)$ is regular, in the same sense as in Proposition 2. Suppose the orthonormal functions $r_k(\cdot), k = 1, \ldots, d$, are continuous and bounded on $[0,1]^p$. Then*

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

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

where r_k are realised according to (19)*;*

(ii) for the empirical regression process, based on residuals $\hat{e}$ of (16)*,*

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

the following convergence of the covariance function is true:

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

where $Q_k(x) = \int_{z \le x} r_(z)dz$;
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 $r_k(\cdot), k = 1, \ldots, d$.

Given two orthonormal systems of n-vectors $(\mu_{\theta,k})_{k=1}^d$) and $(r_k)_{k=1}^d$ the operator K_d, defined in the lemma, will rotate one system into another, regardless of how these systems have been constructed. Therefore (17) is again true for p-dimensional time, and it implies (i).

To see that (ii) is true, one can follow the proof of (ii) in Proposition 2 using (20) for $\mathbb{I}_x^T \mathbb{I}_y$ and using (21) for $\frac{1}{\sqrt{n}} \mathbb{I}_x^T r_k$.

Few more comments and illustrations to follow.

References

[Bates, Watts (2007)] Douglas M. Bates, Donald G . Watts, *Nonlinear Regression Analysis and Its Applications,* Second edition, Wiley (2007)

[Cook, Weisberg (1982)] R. Dennis Cook, Sanford Weisberg, *Residuals and influence in regression,* Chapman and Hall, New York, 1982

[del Bario *et al* (2018)] Eustasio del Barrio, Juan A. Cuesta-Albertos, Marc Hallin, Carlos Matrán, Center-Outward Distribution Functions, Quantiles, Ranks, and Signs in $\mathbb{R}^d$, 2018, arXiv:1806.01238

[Chown, Müller (2018)] Justin Chown and Ursula U. Müller Detecting heteroscedasticity in non-parametric regression using weighted empirical processes, J.R. Statist. Soc., B, 80 (2018), 961-974

[Dette, Hetzler (2009)] Dette, H. and Hetzler, B. A simple test for the nonparametric form of the variance function in nonparametric regression. Annals Institute Statistical Mathematics, 61 (2009), 861886.

[Dette, Munk (1998)] Dette, H. and Munk, A. (1998) Testing heteroscedasticity in nonparametric regression. J. R. Statist. Soc. B, 60 (1998), 693708.

[Dette *et al* (2007)] Dette, H., Neumeyer, N. and Van Keilegom, I. (2007) A new test for the parametric form of the variance function in non-parametric regression. J. R. Statist. Soc. B, 69, 903917.

[Gonzales Manteiga, Crujeiras (2013)] Gonzlez-Manteiga, W. and Crujeiras, R., An updated review of Goodness-of-Fit tests for regression models. TEST 22 (2013), 361-447

[Harrell (2015)] Frank E. Harrell, Jr., *Regression Modelling Strategies,* 2nd ed, Springer, 2015

[Khmaladze(1993)] E.V. Khmaladze, Goodness of fit problems and scanning innovation martinagles, The Annals of Statistics, 21 (1993), 798-829

[Khmaladze and Koul(2004)] Estate V. Khmaladze and Hira L. Koul, Martingale transforms goodness of fit tests in regression models, The Annals of Statistics, 32 (2004), 955-1034

[Khmaladze(2013)] Estate V. Khmaladze, Note on distribution free testing for discrete distributions, The Annals of Statistics, 41 (2013), 2979-2993

[Khmaladze(2016)] Estate Khmaladze, Unitary transformations, empirical processes and distribution free testing, Bernoulli, 22 (2016), 563-588

[Khmaladze(2017)] Estate Khmaladze, Distribution free testing for conditional distributions given covariates, Statistics and Probability Letters, 129 (2017), 348-354

[Keonker(2005)] Roger Koenker, *Quantile Regression*, Cambridge University Press (2005)

[Koul *et al* (2017)] Hira L. Koul, Ursula U. Müller and Anton Schick, Estimating the Error Distribution in a Single-Index Model, Festschrift in Honour of Winfried Stute, *Editors:* Dietmar Ferger, Wenceslao Gonzlez Manteiga, Thorsten Schmidt, Jane-Ling Wang, Springer, 2017

[McCullagh, Nelder(2008)] P. McCullagh, John A. Nelder, *An Introduction to Generalized Linear Models,* Third edition, Chapman & Hall/ CRC Monographs on Statistics (2008)

[Nguyen(2017)] Thuong T.M. Nguyen, New approach to distribution free tests in contingency tables, Metrika, 80 (2017), 153-170

[Müller *et al* (2009)] Ursula U. Müller, Anton Schick, Wolfgang Wefelmeyer, Estimating the innovation distribution in nonparametric autoregression, Probab. Theory Relat. Fields,144 (2009), 5377

[Peyré, Cuturi (2019)] Gabriel Peyté, Marco Cuturi, Computational Optimal Transport, arXiv:1803.00567v2 [statML] 6Feb 2019

[Roberts(2019)] Leigh Roberts, On distribution free goodness of fit testing of Bernoulli trials, Statistics and Probability Letters, 150 (2019), 47-53

[Rosenblatt(1952)] Murray Rosenblatt, Remarks on a Multivariate Transformation, The Annals of Mathematical Statistics, 23 (1952), 470-472

[Segers(2018)] Cees de Valk, Johan Segers, Stability and tail limits of transport-based quantile contours, 2018, arXiv:1811.12061

[Stute(1997)] W. Stute, Nonparametric model checks for regression, The Annals of Statistics, 25 (1997), 613-641

[Villani(2009)] Cedric Villani, *Optimal transport: old and new,* Springer (2009)