



HAL
open science

Empirical L2-distance between regression model and vector space

Zaher Mohdeb, Abdelkader Mokkaemr

► **To cite this version:**

Zaher Mohdeb, Abdelkader Mokkaemr. Empirical L2-distance between regression model and vector space. *Annales de l'ISUP*, 2019, 63 (2-3), pp.166-172. hal-03603880

HAL Id: hal-03603880

<https://hal.science/hal-03603880>

Submitted on 10 Mar 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Pub. Inst. Stat. Univ. Paris

63, fasc. 2-3, 2019, 166-172

Numéro spécial en l'honneur des 80 ans de Denis Bosq /

Special issue in honour of Denis Bosq's 80th birthday

Empirical L^2 -distance between regression model and vector space

Zaher Mohdeb^{†,*} and Abdelkader Mokkadem[‡]

École Nationale Polytechnique de Constantine [†] and Université de Versailles Saint-Quentin-en-Yvelines[‡]

Abstract: The paper is devoted to the estimation of the square distance between the regression function f and the subspace E_p spanned by the linearly independent functions g_1, \dots, g_p . In order to estimate this measure of discrepancy, we propose an empirical L^2 -distance between f and E_p , without weight and we show that it is invariant with respect to change of basis in E_p .

1. Introduction

We consider the following regression model

$$(1.1) \quad Y_{i,n} = f(t_{i,n}) + \varepsilon_{i,n}, \quad i = 1, \dots, n,$$

where f is a unknown real function, defined on the interval $[0, 1]$ and $t_{1,n} = 0 < t_{2,n} < \dots < t_{n,n} = 1$, is a fixed sampling of the interval $[0, 1]$. The errors $\varepsilon_{i,n}$ form a triangular array of random variables with expectation zero and finite variance σ^2 , and for any n , the random variables $\varepsilon_{1,n}, \dots, \varepsilon_{n,n}$ are independent.

Let g_1, \dots, g_p be linearly independent functions defined on $[0, 1]$ and let E_p be the vector space spanned by g_1, \dots, g_p . We assume that the design $\{t_{1,n}, \dots, t_{n,n}\}$ is associated to a positive density function h in the interval $[0, 1]$. We denote $L^2(d\mu)$, where $d\mu(t) = h(t) dt$, the space of square integrable functions, equipped with the usual inner product $\langle \cdot, \cdot \rangle$. As measure of discrepancy between the regression function f and the subspace E_p , we may use the square distance

$$(1.2) \quad \mathcal{D}^2(f) = \min_{v \in E_p} \int_0^1 |f(t) - v(t)|^2 h(t) dt.$$

In this paper, we propose an empirical L^2 -distance of $\mathcal{D}^2(f)$. We follow the procedure of Dette and Munk (1998), but without use of weights. Indeed Dette and Munk (1998) use weights satisfying particular conditions; this leads to a problem with the choice of weights. Generally, this empirical L^2 -distance is used to test the null

*Laboratoire de Mathématiques et Sciences de la Décision, Université frères Mentouri de Constantine, Algérie

AMS 2000 subject classifications: Primary 62G08, 62G05; secondary 62J02

Keywords and phrases: Regression model, Empirical distance

hypothesis H_0 : " $f \in E_p$ " against the alternative hypothesis H_1 : " $f \notin E_p$ ". The previous test has been the subject of several works: Dette and Munk (1998), Mohdeb and Mokeddem (2002, 2004) introduce a test statistic based on the estimation of the square of the distance of f to E_p with respect to a L^2 -norm. Other related work includes that of Dette (1999), González-Manteiga and Cao (1993) and Stute (1997). We give the main result in the following section.

2. Main result

First, note that square distance $\mathcal{D}^2(f)$ given in (1.2) may be expressed also as

$$(2.1) \quad \mathcal{D}^2(f) = \frac{G(f, g_1, \dots, g_p)}{G(g_1, \dots, g_p)},$$

where $G(v_1, \dots, v_k)$ denotes the Gramian determinant $|\langle v_i, v_j \rangle|_{i,j=1,\dots,k}$ for v_1, \dots, v_k in $L^2(d\mu)$. We need to estimate $\mathcal{D}^2(f)$; for this, we use the observations $Y = (Y_{1,n}, \dots, Y_{n,n})'$ and we follow the procedure of Dette and Munk (1998), but without use of weights.

Let

$$\Delta_{i,n} = t_{i,n} - t_{i-1,n}, \quad i = 2, \dots, n, \quad \Delta_{1,n} = \Delta_{2,n},$$

$$W = \text{diag} \left(\Delta_{i,n} h(t_{i,n}) \right)_{i=1,\dots,n}$$

and let $E_{p,n}$ be the vector space of \mathbb{R}^n spanned by $(g_{1,n}, \dots, g_{n,n})$, where

$$g_{k,n} = (g_k(t_{1,n}), \dots, g_k(t_{n,n}))', \quad k = 1, \dots, n.$$

Dette and Munk (1998) propose an estimator denoted \hat{M}^2 of the distance $\mathcal{D}^2(f)$ between f and the model E_p , and estimating the unknown inner products involved in the expression (2.1) by empirical expressions. To show the invariance of the estimator \hat{M}^2 with respect to change of the basis in E_p , Dette and Munk (1998) are led to introduce suitable weights $w_{i,n}$, i, \dots, n satisfying the following particular conditions:

$$\sum_{i=1}^n \Delta_{i,n} w_{i,n} h(t_{i,n}) = 1 \quad \forall n \geq 1$$

and

$$\langle g_k, g_l \rangle = \sum_{i=1}^n \Delta_{i,n} w_{i,n} h(t_{i,n}) g_k(t_{i,n}) g_l(t_{i,n}), \quad 1 \leq k \leq l \leq p.$$

Weights $w_{i,n}$, i, \dots, n satisfying the previous conditions are generally not easy to obtainable (see the discussion on pages 783-784 in Dette and Munk (1998), and Theorem 2.1 therein).

In this paper, we define $G_n(Y, g_1, \dots, g_p)$ as the determinant obtained by replacing in (2.1) the inner product $\langle f, f \rangle$ by

$$Y'WY = \sum_{i=1}^n \Delta_{i,n} h(t_{i,n}) Y_{i,n}^2$$

and $\langle f, g_k \rangle$ by

$$Y'Wg_{k,n} = \sum_{i=1}^n \Delta_{i,n} h(t_{i,n}) g_k(t_{i,n}) Y_{i,n}, \quad k = 1, \dots, p.$$

We estimate $\mathcal{D}^2(f)$ by

$$\mathcal{D}_n^2 = \frac{G_n(Y, g_1, \dots, g_p)}{G(g_1, \dots, g_p)}.$$

It is clear that $\mathcal{D}^2(f)$ does not depend on the basis $\{g_1, \dots, g_p\}$ of E_p , since $\mathcal{D}^2(f)$ is the square distance in $L^2(d\mu)$ of f to E_p , but it is not obvious the same holds for $\mathcal{D}_n^2(Y, g_1, \dots, g_p)$. The following result states that $\mathcal{D}_n^2(Y, g_1, \dots, g_p)$ does not depend on the choice of the basis $\{g_1, \dots, g_p\}$.

Theorem 2.1. *For any $\Theta \in \mathbb{R}^n$, $\mathcal{D}_n^2(\Theta, g_1, \dots, g_p)$ is invariant with respect to change of basis in E_p .*

Proof. Let $\mathcal{U} = \{u_1, \dots, u_p\}$ be an orthonormal basis of E_p and $\mathcal{V} = \{v_1, \dots, v_p\}$ an arbitrary basis of E_p . It is sufficient to show that

$$\mathcal{D}_n^2(\Theta, u_1, \dots, u_p) = \mathcal{D}_n^2(\Theta, v_1, \dots, v_p).$$

We have for all $f \in L^2(d\mu)$,

$$\mathcal{D}^2(f) = \langle f, f \rangle + P(\langle f, u_1 \rangle, \dots, \langle f, u_p \rangle)$$

and

$$\mathcal{D}^2(f) = \langle f, f \rangle + Q(\langle f, v_1 \rangle, \dots, \langle f, v_p \rangle),$$

where P (resp. Q) is a second degree polynomial whose coefficients depend only of the basis \mathcal{U} (resp. \mathcal{V}).

Let $A = (a_{ij})_{i,j=1,\dots,p}$ be the change-of-basis matrix, that is

$$(2.2) \quad v_k = a_{k1}u_1 + \dots + a_{kp}u_p, \quad k = 1, \dots, p.$$

We have

$$\langle f, v_k \rangle = a_{k1}\langle f, u_1 \rangle + \dots + a_{kp}\langle f, u_p \rangle, \quad k = 1, \dots, p,$$

then

$$\left(\langle f, v_1 \rangle, \dots, \langle f, v_p \rangle\right) = H\left(\langle f, u_1 \rangle, \dots, \langle f, u_p \rangle\right),$$

where $H: \mathbb{R}^p \rightarrow \mathbb{R}^p$ is polynomial map, with $H(c_1, \dots, c_p) = (c_1, \dots, c_p)A'$.

Therefore

$$\mathcal{D}^2(f) = \langle f, f \rangle + Q \circ H\left(\langle f, u_1 \rangle, \dots, \langle f, u_p \rangle\right),$$

whence

$$P\left(\langle f, u_1 \rangle, \dots, \langle f, u_p \rangle\right) = Q \circ H\left(\langle f, u_1 \rangle, \dots, \langle f, u_p \rangle\right), \quad \forall f \in L^2(d\mu).$$

Otherwise, $\forall (z_1, \dots, z_p) \in \mathbb{R}^p$, $\exists f = \sum_{j=1}^p z_j u_j \in L^2(d\mu)$ such that $z_j = \langle f, u_j \rangle$, $j = 1, \dots, p$; therefore

$$P(z_1, \dots, z_p) = Q \circ H(z_1, \dots, z_p), \quad \forall (z_1, \dots, z_p) \in \mathbb{R}^p,$$

which leads to $P \equiv Q \circ H$.

Consider now $\mathcal{D}_n^2(\Theta, u_1, \dots, u_p)$ and $\mathcal{D}_n^2(\Theta, v_1, \dots, v_p)$.

We have

$$\mathcal{D}_n^2(\Theta, u_1, \dots, u_p) = \Theta'W\Theta + P\left(\Theta'Wu_{1,n}, \dots, \Theta'Wu_{p,n}\right)$$

and

$$\mathcal{D}_n^2(\Theta, v_1, \dots, v_p) = \Theta'W\Theta + Q\left(\Theta'Wv_{1,n}, \dots, \Theta'Wv_{p,n}\right),$$

where $W = \text{diag}(\Delta_{i,n}h(t_{i,n}))_{i=1, \dots, n}$,

$$u_{k,n} = (u_k(t_{1,n}), \dots, u_k(t_{n,n}))'$$

and

$$v_{k,n} = (v_k(t_{1,n}), \dots, v_k(t_{n,n}))'.$$

According to (2.2), we have also

$$v_{k,n} = a_{k1}u_{1,n} + \dots + a_{kp}u_{p,n}, \quad k = 1, \dots, p$$

and therefore

$$\Theta'Wv_{k,n} = a_{k1}\Theta'Wu_{1,n} + \dots + a_{kp}\Theta'Wu_{p,n} \quad k = 1, \dots, p,$$

that is

$$\left(\Theta'Wv_{1,n}, \dots, \Theta'Wv_{p,n}\right) = H\left(\Theta'Wu_{1,n}, \dots, \Theta'Wu_{p,n}\right).$$

Thus

$$\mathcal{D}_n^2(\Theta, v_1, \dots, v_p) = \Theta'W\Theta + Q \circ H\left(\Theta'Wu_{1,n}, \dots, \Theta'Wu_{p,n}\right),$$

as $Q \circ H \equiv P$, therefore

$$\begin{aligned} \mathcal{D}_n^2(\Theta, v_1, \dots, v_p) &= \Theta'W\Theta + P\left(\Theta'Wu_{1,n}, \dots, \Theta'Wu_{p,n}\right) \\ &= \mathcal{D}_n^2(\Theta, u_1, \dots, u_p). \end{aligned}$$

The result is proved. \square

References

- [1] DETTE, H. and MUNK, A. (1998). Validation of linear regression models. *Ann. Statist.*, **26**, 778–800.
- [2] DETTE, H. (1999). A consistent test for the functional form of a regression based on a difference of variance estimators. *Ann. Statist.*, **27**, 1012–1040.
- [3] GONZÁLEZ-MANTEIGA, W. and CAO, R. (1993). Testing the hypothesis of a general linear model using nonparametric regression estimation. *Test*, **2**, 161–188.
- [4] MOHDEB, Z. and MOKKADEM, A. (2002). Testing the goodness-of-fit of a linear model in nonparametric regression. Goodness-of-fit tests and model validity (Paris, 2000), *Stat. Ind. Technol., Birkhäuser Boston, Boston, MA*, 185–193.
- [5] MOHDEB, Z. and MOKKADEM, A. (2004). Average squared residuals approach for testing linear hypotheses in nonparametric regression. The International Conference on Recent Trends and Directions in Nonparametric Statistics. *J. Nonparametr. Stat.* **16**, no. 1-2, 3–12.
- [6] STUTE, W. (1997). Nonparametric model checks for regression. *Ann. Statist.*, **25**, 613–641.

École Nationale Polytechnique de Constantine,
Campus de l'Université de Constantine 3, Algérie
and
Laboratoire de Mathématiques
et Sciences de la Décision,
Université frères Mentouri Constantine, Algérie
e-mail: zaher.mohdeb@umc.edu.dz

Université de Versailles
Saint-Quentin-En-Yvelines
Bâtiment Fermat,
Département de Mathématiques,
45, Avenue des Etats-Unis,
78035 Versailles Cedex, France
e-mail: abdelkader.mokkadem@uvsq.fr