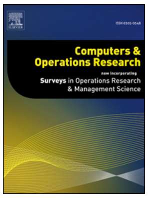
Computers and Operations Research 95 (2018) 172–193

Contents lists available at ScienceDirect

Computers and Operations Research

journal homepage: www.elsevier.com/locate/cor



Locating hyperplanes to fitting set of points: A general framework

Víctor Blanco[a,*], Justo Puerto[b], Román Salmerón[c]

[a] Dpt. Quant. Methods for Economics & Business and IEMath-GR, Universidad de Granada, Spain
[b] Dpt. Statistics & OR and IMUS, Universidad de Sevilla, Spain
[c] Dpt. Quant. Methods for Economics & Business, Universidad de Granada, Spain

ARTICLE INFO

Article history:
Received 17 November 2016
Revised 21 March 2018
Accepted 22 March 2018
Available online 23 March 2018

2010 MSC:
90B85
90C26
52C35
65D10

Keywords:
Fitting hyperplanes
Mathematical programming
Location of structures
Robust fitting
Linear regression

ABSTRACT

This paper presents a family of methods for locating/fitting hyperplanes with respect to a given set of points. We introduce a general framework for a family of aggregation criteria, based on ordered weighted operators, of different distance-based errors. The most popular methods found in the specialized literature, namely least sum of squares, least absolute deviation, least quantile of squares or least trimmed sum of squares among many others, can be cast within this family as particular choices of the errors and the aggregation criteria. Unified mathematical programming formulations for these methods are provided and some interesting cases are analyzed. The most general setting give rise to mixed integer nonlinear programming problems. For those situations we present inner and outer linear approximations to assess tractable solution procedures. It is also proposed a new goodness of fitting index which extends the classical coefficient of determination and allows one to compare different fitting hyperplanes. A series of illustrative examples and extensive computational experiments implemented in R are provided to show the applicability of the proposed methods.

1. Introduction

The problem of locating hyperplanes with respect to a given set of point is well-known in Location Analysis (LA) Schöbel (1999). This problem is closely related to another common question in Data Analysis (DA): to study the behavior of a given set of data with respect to a fitting body expressed with an equation of the form $f(x)=0$, with $x=(X_1,\dots,X_d)\in\mathbb{R}^d$. This last problem reduces to the estimation of the 'best' function f that expresses the relationship between the data or, in the jargon of LA, to the location of the surface $f(x)=0$ that minimizes some aggregation function of the distances to these points (see Amaldi et al., 2016; Diaz-Báñez et al., 2004; Drezner et al., 2002). In many cases the family of functions where f belongs to is fixed and then, the parameters defining such an *optimal* function must be determined. The family of linear functions is the most widely used. This implies that the above equation is of the form $f(x)=\beta_0+\sum_{k=1}^{d}\beta_k X_k=0$ for $\beta_0,\beta_1,\dots,\beta_d\in\mathbb{R}$.

To perform such a fitting, we are given a set of points $\{x_1,\dots,x_n\}\subset\mathbb{R}^d$, and the goal is to find the vector $\hat{\boldsymbol{\beta}}=(\hat{\beta}_0,\hat{\beta}_1,\dots,\hat{\beta}_d)$ that minimizes some measure of the deviation of the data with respect to the hyperplane it induces, $\mathcal{H}(\hat{\boldsymbol{\beta}})=\{z\in\mathbb{R}^d:\hat{\beta}_0+\sum_{k=1}^{d}\hat{\beta}_k z_k=0\}$. For a given point $x\in\mathbb{R}^d$, we define the *residual* with respect to a generic x as a mapping $\varepsilon_x:\mathbb{R}^{d+1}\to\mathbb{R}_+$, that maps any set of coefficients $\boldsymbol{\beta}=(\beta_0,\dots,\beta_d)\in\mathbb{R}^{d+1}$, into a measure $\varepsilon_x(\boldsymbol{\beta})$ that represents the deviation of the given point x from the hyperplane with those parameters. The problem of locating a hyperplane for a given set of points $\{x_1,\dots,x_n\}\subseteq\mathbb{R}^d$ consists of finding the coefficients minimizing an aggregation function, $\Phi:\mathbb{R}^n\to\mathbb{R}$, of the residuals of all the points. Different choices for the residuals and the aggregation criteria will give, in general, different optimal values for the parameters and thus different properties for the resulting hyperplanes. This problem is not new and some of these criteria, as the minisum, minimax and some other alternatives, have been widely analyzed from a LA perspective (see Carrizosa and Plastria, 1995; Megiddo and Tamir, 1983; Schöbel, 1996; Schöbel, 1997; Schöbel, 1998; Schöbel, 1999, among other).

A first approach to locate a hyperplane is to consider that residuals, with respect to given points, are individual measures of error and thus, each residual should be minimized independently of the remaining (Carrizosa et al., 1995; Narula and Wellington, 2007). It is clear that this simultaneous minimization will not be possible in most of the cases and then several strategies can be followed: one can try to find the set of Pareto fitting curves (Carrizosa et al.,

[*] Corresponding author.
E-mail addresses: vblanco@ugr.es (V. Blanco), puerto@us.es (J. Puerto), romansg@ugr.es (R. Salmerón).

https://doi.org/10.1016/j.cor.2018.03.009
0305-0548/© 2018 Elsevier Ltd. All rights reserved.

1995) or alternatively, to apply an aggregation function that incorporates the holistic preference of the Decision-Maker on the different residuals (Yager and Beliakov, 2010). This last choice is very difficult and it is usual to apply an approach of *complete uncertainty* (i.e., it is assumed that it is known the set of possible outcomes, but there is no information about the probabilities of those outcomes or about their likelihood ranking) leading to additive aggregations.

The most popular methods to compute the coefficients of an optimal hyperplane consider that the residuals are the differences from one of the coordinates of the space (which are usually known as vertical/horizontal distances). In this paper we present a framework that generalizes previous contributions for optimally locating/fitting hyperplanes to a set of points. It introduces a family of combinations residuals-criteria that allows for a great flexibility to accommodate hyperplanes to a set of points (Marín et al., 2009; Nickel and Puerto, 2005). One of the contributions of our proposal is the use of modern mathematical programming tools to solve the problems which are involved in the computation of the parameters of the fitting models. In addition, it can be combined with some of the mathematical programming techniques for feature selection (Bertsimas and Mazumder, 2014), with classification schemes (Bertsimas and Shioda, 2007), or with constraints on the coefficients of the linear manifold. This unified framework is also able to accommodate general forms of regularization, as upper bound on the ℓ_2-norm of the coefficients (Hoerl and Kennard, 1988), since it would only mean to add additional constraints to the mathematical programming formulations proposed in the paper, at the price of increasing the computational complexity needed for solving the problems. Many of the formulations described in this paper have been implemented in R in order to be available for data analysts.

In our framework, errors are measured as shortest distances, based on a norm, between the given points and the fitting surface. This makes the location problem geometrically invariant which is an interesting advance with respect to vertical/horizontal residuals. We observe that this framework subsumes as particular cases the standard location methods that consider residuals based on vertical distances (commonly used in Statistics); as well as most of the particular cases of fitting linear bodies using vertical distances but different aggregation criteria described in the literature, as ℓ_p fitting (ℓ_p-norm criterion), least quantile of squares (Bertsimas and Mazumder, 2014; Rousseeuw, 1984), least trimmed sum of squares (Atkinson and Cheng, 1999; Rousseeuw, 1983), etc. The use of nonstandard residuals is common in the area of LA and other areas of Operations Research. However, it is not that usual in the field of regression analysis although orthogonal (ℓ_2) residuals have been already used, see, e.g., Euclidean Fitting (Bargiela and Hartley, 1993; Cavalier and Melloy, 1991; Pinson et al., 2008) or Total Least Squares (Van Huffel and Vanderwalle, 1991), mainly applied to bidimensional data; and the more general geodesic distance residuals are applied in geodesic regression (Fletcher, 2013). Quoting the reasons for that fact given by Giloni and Padberg (2002): "we have left out a summary of linear regression models using the more general ℓ_τ-norms with $\tau \notin \{1, 2, \infty\}$ for which the computational requirements are considerably more burdensome than in the linear programming case (as they generally require methods from convex programming where machine computations are far more limited today)."

In order to compare the *goodness of the fitting* for the different models, we have developed a new generalized measure of fit. This proposal is based on a generalization of the classical coefficient of determination for least squares fitting, that will allow one to measure how good is an optimal hyperplane with respect to the best constant model, $X_d = \beta_0$.

The paper is organized as follows. In Section 2 we introduce the new framework for fitting hyperplanes as well as some results that allow us to interpret the obtained solutions for practical purposes. Next, in Section 3, a residual-aggregation dependent goodness of fitting index is defined and an efficient approach for its computation is presented. Two types of residuals are analyzed in more detail, namely those induced by polyhedral-and-ℓ_τ norms for rational $\tau \geq 1$. In Section 4, we present new methods for the location of hyperplanes assuming that the residuals are measured as the shortest norm-based distance between the given points (data set) and the linear fitting body using polyhedral norms. The results of this section are instrumental. They constitute the basis to address the more general problems in Section 5, since they will permit to develop inner and outer linear approximations for more general Mixed Integer Non Linear Programming (MINLP) problems that result in the general case. Section 5 analyzes the location of hyperplanes using ℓ_τ norms. Since in this case non convex problems are derived, we also present outer and inner linear approximations that reduce, the corresponding MINLP problems with ℓ_τ-norms residuals, to problems with polyhedral norm residuals. Section 6 is devoted to the computational experiments. We report results for synthetic data and for the classical data set given in Durbin and Watson (1951). In addition, we include an illustrative example of the scalability of the methodology with several thousands of points. The paper finishes with some concluding remarks and future research.

2. A flexible methodology for the location of hyperplanes

Given is a set of n observations or demand points (depending that we use the *jargon* of data analysis or location analysis, respectively) in a $(d+1)$-dimensional space, $\{x_1, \ldots, x_n\} \subset \{1\} \times \mathbb{R}^d$ (we will assume, for a clearer description of the models, that the first, the 0th, component of x_i is the one that account for the intercept, being $x_{10} = \cdots = x_{n0} = 1$). Next, we analyze the problem of locating a linear form (hyperplane) to fit these points minimizing different forms of measuring the residuals and their aggregation. For any $y \in \mathbb{R}^{d+1}$, we shall denote $y_{-0} = (y_1, \ldots, y_d)$, i.e., the vector with the last d coordinates of y excluding the first one. First, we assume that the point-to-hyperplane deviation is modeled by a residual mapping $\varepsilon_x : \mathbb{R}^{d+1} \to \mathbb{R}_+$, $\varepsilon_x(\boldsymbol{\beta}) = \mathrm{D}(x_{-0}, \mathcal{H}(\boldsymbol{\beta}))$, being D a distance measure in $\mathbb{R}^d$. This residual represents how "far" is the point (observation) $x \in \mathbb{R}^{d+1}$ with respect to the hyperplane $\mathcal{H}(\boldsymbol{\beta}) = \{y \in \mathbb{R}^d : (1, y^t)\boldsymbol{\beta} = 0\}$. At times, for the sake of brevity, we will write the hyperplane as $\boldsymbol{\beta}^t X = 0$, with $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_d)^t \in \mathbb{R}^{d+1}$. In addition, to simplify the presentation, we will refer, whenever no possible confusion occurs, to the residual with respect to the point x_i as ε_i.

An overall measure of the deviations of the whole data set with respect to the hyperplane induced by $\boldsymbol{\beta}$ is obtained by using an aggregation function of the residuals, $\Phi : \mathbb{R}^n \to \mathbb{R}$. With this setting, one tries to minimize such an aggregation function and the *Fitting Hyperplane Problem* (FHP) consists of finding $\hat{\boldsymbol{\beta}} \in \mathbb{R}^{d+1}$ such that:

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^{d+1}} \Phi(\boldsymbol{\varepsilon}(\boldsymbol{\beta})), \tag{1}$$

where $\boldsymbol{\varepsilon}(\boldsymbol{\beta}) = (\varepsilon_1(\boldsymbol{\beta}), \ldots, \varepsilon_n(\boldsymbol{\beta}))^t$ is the vector of residuals.

Note that the difficulty of solving Problem (1) depends on both the expressions for the residuals and the aggregation criterion Φ. If Φ and ε_x are linear, the above problem becomes a linear programming problem. In this paper, we consider a general family of aggregation criteria that includes as particular cases most of the classical ones used in the literature (Bertsimas and Mazumder, 2014; Giloni and Padberg, 2002; Rousseeuw and Leroy, 2003; Yager and Beliakov, 2010).

Let $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$ and let $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ be the vector of residuals of all of the points in the given data set. We consider aggregation criteria

$\Phi : \mathbb{R}^n \to \mathbb{R}_+$ defined as:

$$\Phi(\boldsymbol{\varepsilon}) = \sum_{i=1}^{n} \lambda_i \, \boldsymbol{\varepsilon}^p_{(i)}, \quad 1 \le p < +\infty, \tag{2}$$

where $\boldsymbol{\varepsilon}_{(i)} \in \{\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n\}$ is such that $\boldsymbol{\varepsilon}_{(1)} \le \cdots \le \boldsymbol{\varepsilon}_{(n)}$. Observe that this operator defines a multiparametric family (called *ordered median functions* (Nickel and Puerto, 2005)) that depending on the choice of the λ-weights captures many of the models proposed in the literature.

Most classical models assume that the residuals are defined as the vertical distance (with respect to the last coordinate) from the points to the hyperplane:

$$\boldsymbol{\varepsilon}_x(\boldsymbol{\beta}) = \left| x_d - \sum_{k=0}^{d-1} \frac{\beta_k}{\beta_d} x_k \right|, \tag{3}$$

(assuming that $\beta_d \neq 0$).

Therefore, the difference between them comes from the choice of the aggregation criterion Φ. We show below how some classical methods can be accommodated to our framework.

1. The Least Sum of Squares (LSS) method, credited to Gauss (1809), is the most widely used approach to estimate the coefficients of a linear model due to its simplicity (a closed form for the optimal coefficients is obtained) and its theoretical implications for the inference over the total population. However, somehow restricting hypotheses are required in order to be applied (see, e.g., Giloni and Padberg, 2002). The LSS criterion is defined as the sum of the squares of the residuals, that is: $\Phi_{LSS}(\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n) = \sum_{i=1}^n \boldsymbol{\varepsilon}_i^2$, where the residuals $\boldsymbol{\varepsilon}_k$ are given by (3). The reader may observe that LSS corresponds to Problem (1) with $\lambda^t = (1, \ldots, 1)$, $p = 2$ and $\boldsymbol{\varepsilon}$ the vertical distance.
2. The Least Absolute Deviation (LAD) method (introduced by Edgeworth, 1887) consists of minimizing the sum of the absolute value of the vertical residuals. Therefore, $\Phi_{LAD}(\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n) = \sum_{i=1}^n |\boldsymbol{\varepsilon}_i|$. Note that LAD corresponds to the model in (1) for $\lambda^t = (1, \ldots, 1)$ and $p = 1$.
3. The Least Quantile of Squares (LQS), recently introduced by Bertsimas and Mazumder (2014), is a generalization of the Least Median of Squares (LMS) introduced by Hampel (1975). It also considers vertical distances as residuals, but they are aggregated to minimize the r-quantile of its distribution (r ranges in $\{1, \ldots, n\}$). Hence, $\Phi_{LQS}(\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n) = r-$quantile$(\boldsymbol{\varepsilon}_1^2, \ldots, \boldsymbol{\varepsilon}_n^2) := \boldsymbol{\varepsilon}^2_{(r)}$.
   This method also fits to the general form of the aggregation criteria considered in this paper. In this case, following the notation introduced in (2), the LQS hyperplane can be obtained for $p = 2$ and $\lambda = (\overbrace{0, \ldots, 0}^{(r-1)}, 1, \overbrace{0, \ldots, 0}^{(n-r)})$. (Observe that LMS hyperplane is also obtained within the same scheme when $p = 2$ and $\lambda = (\overbrace{0, \ldots, 0}^{\lfloor \frac{n}{2} \rfloor},\ 1\ , \overbrace{0, \ldots, 0}^{\lfloor \frac{n}{2} \rfloor})$.)
4. The Least Trimmed Sum of Squares (LTS) method was introduced by Rousseeuw (1984) as a robust alternative to the LSS method, in that it has a high breakdown point. Recall that, intuitively, the *breakdown point* of an estimator is the proportion of incorrect observations (e.g., arbitrarily large observations) an estimator can handle before giving an incorrect (e.g., arbitrarily large) result. With our notation, it corresponds to choose again as residuals the vertical distance, $p = 2$, and the aggregation criterion $\Phi_{LTS}(\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n) = \sum_{i=1}^h \boldsymbol{\varepsilon}^2_{(i)}$ where $\boldsymbol{\varepsilon}_{(i)} \in \{\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n\}$ with $\boldsymbol{\varepsilon}_{(i)} \le \boldsymbol{\varepsilon}_{(i+1)}$ for $i = 1, \ldots, n-1$, and $h \in \{1, \ldots, n\}$. The most common choice for h is $\lfloor \frac{n}{2} \rfloor$, considering the best 50% square residuals.
   In the following, we denote by $LTS(\alpha)$ the LTS method when $100 - \alpha\%$ of the data is discarded, i.e., the percentage of the data that may be considered as outliers.

The function Φ, introduced in (2), is invariant against permutations of its components (sometimes called *symmetric* in the related literature) and, for non negative lambda weights, a monotone function, ensuring that the ordering of the individual residuals do not affect the overall goodness of the fitting. Moreover, it also implies that a componentwise smaller vector of residuals gives rise to a more accurate fitting.

The natural implication of the assumption made about the definition of residuals is that, as expected, the response (projection) of a point on a given hyperplane differs from the classical evaluation. In this setting the response is the closest point, with respect to the distance D, to the hyperplane $\mathcal{H}(\boldsymbol{\beta})$. For the sake of readability, we include the following result which follows applying (Mangasarian, 1999, Theorem 2.1) to the definition of the residual mapping $\boldsymbol{\varepsilon}_z = \min_{y \in \mathcal{H}(\boldsymbol{\beta})} \|z_{-0} - y\|$.

Lemma 2.1. *For a given point* $z^t = (1, z_1, \ldots, z_d)$ *and the hyperplane* $\mathcal{H}(\boldsymbol{\beta})$ *the response* $\hat{z}$ *consistent with the residual* $\boldsymbol{\varepsilon}_z = \min_{y \in \mathcal{H}(\boldsymbol{\beta})} \|z_{-0} - y\|$ *is given by* $\hat{z} = z_{-0} - \frac{\boldsymbol{\beta}^t z}{\|\boldsymbol{\beta}_{-0}\|^*} \mathrm{k}(\boldsymbol{\beta})$, *where* $\|y\|^* = \max_{z \in \mathbb{R}^d : \|z\| \le 1} z^t y$ *is the dual norm to* $\|y\|$ *and* $\mathrm{k}(\boldsymbol{\beta}) = \arg\ \max_{\|x\|=1} \boldsymbol{\beta}^t_{-0} x$. *Moreover,*

$$\boldsymbol{\varepsilon}_z = \frac{|\boldsymbol{\beta}^t z|}{\|\boldsymbol{\beta}_{-0}\|^*}. \tag{4}$$

From the above result, the response for a point with a unknown coordinate (without loss of generality, the last component, d), namely $z = (1, z_1, \ldots, z_{d-1}, 0)^t$, will be given by:

$$\hat{z}_d = -\frac{\boldsymbol{\beta}^t z}{\|\boldsymbol{\beta}_{-0}\|^*} \mathrm{k}(\boldsymbol{\beta})_d.$$

Hence, differentiating $\hat{z}$ with respect to each z_j, $j = 1, \ldots, d-1$, we get

$$\frac{\partial \hat{z}_d}{\partial z_j} = -\frac{\beta_j}{\|\boldsymbol{\beta}_{-0}\|^*} \mathrm{k}(\boldsymbol{\beta})_d,$$

which may be interpreted as the marginal variation of the d-th coordinate with respect to the jth coordinate whenever the other dimensions remain constant.

Explicit expressions for such projections, namely, ℓ_1, ℓ_∞ and ℓ_τ-norms, for $\tau > 1$ are described in the following lemma.

Lemma 2.2. *Let* $z = (1, z_1, \ldots, z_d)^t$, *then*

1. *If* D *is the* ℓ_1*- distance,*
$$\hat{z}_k = \begin{cases} z_k & \text{if } |\beta_k| \neq \max\{|\beta_j| : j = 1, \ldots, d\}, \\ z_k - \dfrac{\boldsymbol{\beta}^t z}{\|\boldsymbol{\beta}_{-0}\|_\infty} \nu_k, & \text{if } \beta_k = \max\{|\beta_j| : j = 1, \ldots, d\}, \\ z_k + \dfrac{\boldsymbol{\beta}^t z}{\|\boldsymbol{\beta}_{-0}\|_\infty} \nu_k, & \text{if } \beta_k = -\max\{|\beta_j| : j = 1, \ldots, d\}, \end{cases}$$
for $k = 1, \ldots, d$, *and for some* $\nu_1, \ldots, \nu_d \ge 0$ *such that* $\sum_j \nu_j = 1$.
2. *If* D *is the* ℓ_∞*- distance,*
$$\hat{z}_k = \begin{cases} z_k - \dfrac{\boldsymbol{\beta}^t z}{\|\boldsymbol{\beta}_{-0}\|_1}, & \text{if } \beta_k > 0, \\ z_k + \dfrac{\boldsymbol{\beta}^t z}{\|\boldsymbol{\beta}_{-0}\|_1}, & \text{if } \beta_k < 0, \end{cases} \qquad k = 1, \ldots, d.$$
3. *If* D *is the* ℓ_τ*- distance with* $1 < \tau < +\infty$ *then*
$$\hat{z}_k = z_k - \frac{\boldsymbol{\beta}^t z}{\|\boldsymbol{\beta}_{-0}\|_\nu} \mathrm{k}_\tau(\boldsymbol{\beta})_k, \quad k = 1, \ldots, d$$

and

$$\mathrm{k}_\tau(\boldsymbol{\beta})_k = \begin{cases} \dfrac{\operatorname{sign}(\boldsymbol{\beta}_k)|\boldsymbol{\beta}_k|^{\nu/\tau}}{\left(\displaystyle\sum_{j=1}^{d} |\boldsymbol{\beta}_j|^{\nu}\right)^{1/\tau}} & \text{if } \boldsymbol{\beta}_k \neq 0 \\ 0 & \text{if } \boldsymbol{\beta}_k = 0, \end{cases} \qquad k = 1, \ldots, d,$$

being ν *such that* $\frac{1}{\tau} + \frac{1}{\nu} = 1$.

Proof. The proof of items 1. and 2. can be found in Mangasarian (1999). The proof of item 3. follows from the Lagrangian optimality condition applied to $\max_{\|z\|_\tau = 1} \boldsymbol{\beta}_{-0} z$. First, we observe that a Lagrange multiplier exists since the problem is regular at any point of the ℓ_τ unit ball (Note that the gradient of the unique constraint is always linearly independent.). Next, the Lagrangian function is $\mathrm{L}(z, \lambda) = \boldsymbol{\beta}_{-0} z - \lambda \sum_{k=1}^{d} |z_k|^\tau$. Therefore, its partial derivatives are: $\frac{\partial \mathrm{L}}{\partial z_k} = \beta_k - \lambda \tau |z_k|^{\tau-1} \operatorname{sign}(z_k)$, for all $k = 1, \ldots, d$. Hence, equating to zero the partial derivatives, it follows that for any index k such that $z_k^* \neq 0$

$$\lambda^* = \frac{\beta_k}{\tau |z_k^*|^{\tau-1}} \operatorname{sign}(z_k^*). \tag{5}$$

Let us define the sets $I = \{k : \beta_k > 0\}$, $J = \{k : \beta_k < 0\}$, $K = \{k : \beta_k = 0\}$. Now from Eq. (5), and taking into account that $\|z\|_\tau = 1$, we obtain:

$$|z_k^*|^\tau = \begin{cases} \dfrac{\left(\operatorname{sign}(z_k^*)\beta_k\right)^\nu}{\left(\sum_{j=1}^{d} \operatorname{sign}(z_j^*)\beta_j\right)^\nu} & \text{if } k \in I \cup J, \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, the Hessian of L is diagonal and all its entries are negative, namely $\frac{\partial^2 \mathrm{L}}{\partial z_k^2} = -\lambda \tau (\tau - 1) |z_k^*|^{\tau-2}$. This implies that z^* and λ^* are local maxima.

In the particular case of $\tau = 2$, one can check that $k_2(\boldsymbol{\beta})_k = \beta_k$ which simplifies the above expression.

□

We note in passing that $\boldsymbol{\varepsilon}_x = \mathrm{D}_{\|\cdot\|}(x_{-0}, \mathcal{H}(\boldsymbol{\beta}))$ and thus, according to Lemma 2.1

$$\mathrm{D}_{\|\cdot\|}(x_{-0}, \mathcal{H}) = \frac{|\boldsymbol{\beta}^t x|}{\|\boldsymbol{\beta}_{-0}\|_*}. \tag{6}$$

Observe also that when the points in the data set lie exactly on a hyperplane, $\mathcal{H}$, this hyperplane is always optimal for all versions of Problem (1), although for some specific choices of λ the solution may not be unique and different hyperplanes may be alternative optima.

Remark that the standard residual (vertical distance) is a distance measure that is not induced by a norm, but its expression can be written in an analogous form and so it fits to the shape of the distances that are considered in this paper. In particular, the vertical distance (with respect to the last coordinate) may be defined as $\mathrm{D}_V(x, H) = |\beta_d x_d - \sum_{i=1}^{d-1} \beta_i x_i - \beta_0| / |\beta_d|$.

The above aggregation criteria (2) and residual functions (4) are rather general and exhibit good structural properties. On the one hand, they accommodate most of the already considered fitting methods in the literature. On the other hand, one can always exploit its properties and different representations in order to solve Problem (1). In the following we prove some structural properties that imply the possibility of applying different methodologies to solve (1).

We note, without proof (it can be found in an extended version of this paper (Blanco et al., 2016)), that our globalizing criterion $\Phi(\boldsymbol{\varepsilon}_x(\cdot))$ is a difference of convex (D.C.) functions. This fact allows one to apply all the available results on the optimization of this class of functions (see, e.g., Thoai, 1999). Alternatively, we can give a more efficient representation that helps latter in the resolution of the problem. This representation is based on simpler functions which replace φ by more friendly classes of functions (with regards to the optimization phase) and that permit to get a manageable form of a mathematical program. In the following we include a first mathematical programming formulation for the generalized fitting Problem (1), for any choice of Φ and $\boldsymbol{\varepsilon}_x$.

Theorem 2.3. *Let* $\{x_1, \ldots, x_n\} \subseteq \mathbb{R}^{d+1}$ *be a set of points,* $\lambda \in \mathbb{R}^n_+$, $\Delta_k = \lambda_k - \lambda_{k-1}$, *for* $k = 2, \ldots, n$, $p = \frac{r}{s} \in \mathbb{Q}$ *and* $\|\cdot\|$ *a norm in* $\mathbb{R}^d$. *Problem (1) is equivalent to the following mathematical programming problem:*

$$\min \ \lambda_1 \sum_{i=1}^{n} z_i + \left\{ \sum_{k:\Delta_k > 0} \Delta_k \left((n-k+1) t_k + \sum_{i=1}^{n} z_{ik} \right) + \sum_{k:\Delta_k < 0} (\Delta_k) \sum_{i=1}^{n} \omega_{ik} \right\} \tag{7}$$

$$\text{s.t. } \boldsymbol{\varepsilon}_i \geq \frac{|\boldsymbol{\beta}^t x_i|}{\|\boldsymbol{\beta}_{-0}\|_*}, \quad \forall i = 1, \ldots, n, \tag{8}$$

$$z_i^s \geq \boldsymbol{\varepsilon}_i^r, \quad \forall i = 1, \ldots, n, \tag{9}$$

$$t_k + z_{ik} \geq z_i, \quad i = 1, \ldots, n, \ k = 2, \ldots, n, \ \Delta_k > 0 \tag{10}$$

$$\sum_{i=1}^{n} \gamma_{ik} = n - k + 1, \quad k = 2, \ldots, n : \ \Delta_k < 0 \tag{11}$$

$$\omega_{ik} \leq M \gamma_{ik}, \quad i = 1, \ldots, n, k = 2, \ldots, n : \ \Delta_k < 0 \tag{12}$$

$$\omega_{ik} \leq z_i, \ i = 1, \ldots, n, k = 2, \ldots, n : \ \Delta_k < 0 \tag{13}$$

$$\gamma_{ik} \in \{0, 1\}, \omega_{ik} \geq 0, \ \Delta_k < 0,$$

$$z_{ik}, \ t_k \geq 0, \quad i, k = 1, \ldots, n, \ \Delta_k > 0$$

$$\boldsymbol{\beta} \in \mathbb{R}^{d+1}, \ \varepsilon_i \geq 0, \ i = 1, \ldots, n,$$

where $M > 0$ *is a suitable large constant.*

Proof. Applying the result in Grzybowski et al. (2011, Theorem 3.6) the aggregation function Φ can be equivalently written as

$$\Phi(\boldsymbol{\varepsilon}(\boldsymbol{\beta})) = \lambda_1 \sum_{i=1}^{n} \boldsymbol{\varepsilon}_i(\boldsymbol{\beta})^p + \sum_{k=2}^{n} \Delta_k \theta_k(\boldsymbol{\beta}), \tag{14}$$

where $\theta_k(\boldsymbol{\beta}) = \max\{\boldsymbol{\varepsilon}_{i_1}(\boldsymbol{\beta})^p + \ldots + \boldsymbol{\varepsilon}_{i_{n-k+1}}(\boldsymbol{\beta})^p : \text{ for all } \{i_1, \ldots, i_{n-k+1}\} \subset \{1, \ldots, n\} \text{ such that } i_1 < i_2 < \ldots < i_{n-k+1}\}$. (The reader may observe that the functions θ_k are usually called $(n-k+1)$-centrum in the specialized literature of optimization Nickel and Puerto, 2005.) The z-variables in the formulation represent the residuals raised to the power of $p = \frac{r}{s}$. The objective function (7) has three terms. The first one corresponds to the first one in (14). The terms $(n-k+1)t_k + \sum_{i=1}^{n} z_{ik}$ together with the constraints (10) provide valid representations for the $(n-k+1)$-centrum functions of the elements of the vector $z = (z_1, \ldots, z_n)^t$ whenever Δ_k is positive. On the other hand, if Δ_k is negative the expression $\sum_{i=1}^{n} \omega_{ik}$ together with (12), (13) and $\gamma_{ik} \in \{0, 1\}$ give a valid representation for the $(n-k+1)$-centrum functions of the elements of the vector $z = (z_1, \ldots, z_n)^t$. Finally, (8) and (9) ensure that $z_i = \varepsilon_i^p$, for all $i = 1, \ldots, n$ in the optimal solution of the problem. □

Note that the above problem is a MINLP problem, whose continuous relaxation is in general non convex due to the set of constraints (8). Apart from the mathematical programming formulation above, one may use alternative (in some cases better) formulations for the ordering problems as those provided in Fernández et al. (2014). In particular, some important special ordered median aggregation criteria permit to have a simpler formulation that avoids the use of binary variables. The following result shows a better formulation for the fitting problem under the assumption that $0 \leq \lambda_1 \leq \ldots \leq \lambda_n$. We call this setting for lambda the *monotone case*.

Proposition 2.4. *Let* $\{x_1, \ldots, x_n\} \subset \mathbb{R}^{d+1}$ *be a set of demand points,* $\lambda \in \mathbb{R}^n$*, such that* $0 \leq \lambda_1 \leq \cdots \leq \lambda_n$, $p = \frac{r}{s} \in \mathbb{Q}$ *with* $r > s \in \mathbb{N}$, $\gcd(r, s) = 1$ *and* $\|\cdot\|$ *a norm in* $\mathbb{R}^d$*. Then, Problem (1) is equivalent to the following mathematical programming problem:*

$$\begin{aligned}
\min\ & \sum_{j=1}^{n} v_j + \sum_{i=1}^{n} w_i \\
\text{s.t. }& (8), (9), \\
& v_j + w_i \geq \lambda_i z_j, \forall i, j = 1, \ldots, n, \\
& z_i, \theta_i \geq 0, v, w \in \mathbb{R}^n, \boldsymbol{\beta} \in \mathbb{R}^{d+1}.
\end{aligned}$$

Proof. The proof follows by the representation of the ordering between the residuals by permutation variables, which for $0 \leq \lambda_1 \leq \cdots \leq \lambda_n$, allows one to write the objective function in Problem (1) as an assignment problem which is totally unimodular. Therefore, it can be equivalently rewritten using its dual problem. The interested reader is refereed to Blanco et al. (2014) for further details on this transformation. □

The reader may observe that the nonlinear constraints $z_i^s \geq \boldsymbol{\varepsilon}_i^r$ for all $i = 1, \ldots, n$ can be transformed into a set of second order cone constraints using a simplified version of Lemma 1 in Blanco et al. (2014). This implies that those constraints can be efficiently handled by nowadays nonlinear solvers since they are convex and friendly for the optimization.

Remark 2.5. Let $r, s \in \mathbb{N} \setminus \{0\}$ with $\gcd(r, s) = 1$, and $k = \lfloor \log_2(r) \rfloor$. Then, there exist variables $u_1, \ldots, u_{k-1} \geq 0$ such that each constraint $z^s \geq \boldsymbol{\varepsilon}^r$ in (8) can be equivalently written as constraints in the form: $u_j^2 \leq u_l^{a_j} z^{b_j} \boldsymbol{\varepsilon}^{c_j}$, $\boldsymbol{\varepsilon}^2 \leq u_h u_{h-1}^{d_h} z^{f_h} \boldsymbol{\varepsilon}^{g_h}$, $u_j \geq 0$, with $j = 1, \ldots, k-1$ and such that $1 \leq a_j + b_j + c_j \leq 2$ for given $a_j, b_j, c_j \in \mathbb{Z}_+$ and $d_h, f_h, g_h \in \mathbb{Z}_+$ such that $d_h + b_h + c_h = 1$.

By the above remark, the nonlinear constraints in the form $z^s \geq \boldsymbol{\varepsilon}^r$ are written as second order cone constraints in the form $X^2 \leq YZ$ or $X^2 \leq Y$ (for some choices of the variables X, Y and Z in our model).

Hence, the difficulty of solving Problem (7)–(13), depends essentially on the choice of the residuals since all except constraints (8) are linear or second order cone constraints which can be efficiently handled with nowadays modern optimization techniques. In the next sections we analyze different choices for the residuals.

Remark 2.6 (Subset Selection and Regularization)**.** In the case where the number of points (n) is much smaller than the dimension of the space (d), it is common in Statistics to compute fitting hyperplanes over a smaller dimension space. The new space is determined by those components that, after projecting, permits a good fitting in a lower dimension space. Several methods have been proposed in the recent literature to perform such a computation. If the dimension of the new space, $q < d$, is given, a constraint in the form $\|\beta_{-0}\|_0 \leq q$ (here $\|\cdot\|_0$ stands for the support function or nuclear norm, i.e., the number of nonzero components of the vector) may be included in the mathematical programming formulation (see Bertsimas et al., 2016; Miller, 2002), which gives rise to the so called Subset Selection Problem. If such a dimension is not known, regularization methods that penalize the number of nonzero elements or the size of β_{-0} can be applied to solve the Feature Selection Problem (see Miyashiro and Takano, 2015). Note that both types of approaches can be incorporated in our models although this will increase its computational complexity.

3. Goodness of fitting

After addressing the problem of locating/fitting a hyperplane with respect to a set of points, we will analyze the goodness of this fitting extending the well-known coefficient of determination, R^2, in Regression Analysis. (Recall that the *coefficient of determination* is the proportion of the variance in the dependent variable that is predictable from the independent variable(s).) For the sake of presentation, we assume that the variable that needs to be analyzed as dependent to the others is the last coordinate X_d, or in other words $Y = X_d$. The *goodness of fitting index*, GoF, is defined as:

$$\mathrm{GoF}_{\Phi, \boldsymbol{\varepsilon}} = 1 - \frac{\Phi^*}{\Phi_0^*},$$

where Φ^* is the optimal value of (1), namely $\Phi(\boldsymbol{\varepsilon}_x(\hat{\boldsymbol{\beta}}))$, and Φ_0^* is the optimal value of Problem (1) when it is additionally required that $\boldsymbol{\beta}$ is in the form $\boldsymbol{\beta} = (\beta_0, \overbrace{0, \ldots, 0}^{d-1}, -1)$, i.e., the hyperplane is forced to be constant ($X_d = \beta_0$). Note that the components $1, \ldots, d-1$ do not appear in the model. Hence, Φ_0^* measures the global error assumed by the best fitting *horizontal* hyperplane; whereas $\mathrm{GoF}_{\Phi,\, \boldsymbol{\varepsilon}}$ measures the improvement of the model that considers all the dimensions with respect to the one that omits all (except one) of them. Observe that this coefficient coincides with the classical coefficient of determination provided that the aggregation criterion is the overall sum and the residuals are the squared vertical distances: in that case $\widehat{\beta}_0 = \bar{x}_{\cdot d}$ (the sample mean of the *dependent* variable). Note that GoF is well defined if $\Phi_0^* \neq 0$.

The GoF clearly verifies one of the important properties of the standard coefficient of determination, $0 \leq \mathrm{GoF}_{\Phi,\, \boldsymbol{\varepsilon}} \leq 1$. Furthermore, one may interpret the coefficient as a measure of how good is the best possible hyperplane under certain criterion and residual choice with respect to the best *horizontal* hyperplane. When GoF is close to 0, it is because $\Phi^* \simeq \Phi_0^*$, so not appreciable improvement is given by the complete model (which considers all the components) with respect to the simple constant model; whenever GoF is close to 1, it means that $\Phi^* \ll \Phi_0^*$, being the proposed model significatively better than the constant model (note that $\mathrm{GoF} = 1$ iff $\Phi^* = 0$, i.e., when the model perfectly fits the demand points). Hence, the closer the GoF to one, the better the fitting; whereas the closer to zero, the better is the constant model with respect to the full model.

Observe that the above definition coincides with some of the alternatives to measure the goodness of fitting for robust approaches to the least sum of squares methodology (see McKean and Sievers, 1987).

To obtain the GoF, apart from solving Problem (1) to get Φ^*, we must also solve the problem:

$$\Phi_0^* = \min_{\beta_0 \in \mathbb{R}} \Phi(\mathrm{D}(x_1, \mathcal{H}_0), \ldots, \mathrm{D}(x_n, \mathcal{H}_0)), \tag{15}$$

where $\mathcal{H}_0 = \{y \in \mathbb{R}^d : y_d = \beta_0\}$ for some $\beta_0 \in \mathbb{R}$.

Lemma 3.1. *Let the residual mapping* $\varepsilon_x : \mathbb{R}^{d+1} \to \mathbb{R}_+$ *be induced by a norm* $\|\cdot\|$*. Then, Problem (15) is equivalent to*

$$\Phi_0^* = \min_{\beta_0 \in \mathbb{R}} \Phi(\kappa_{\boldsymbol{\varepsilon}} |x_{1d} - \beta_0|, \ldots, \kappa_{\boldsymbol{\varepsilon}} |x_{nd} - \beta_0|), \tag{16}$$

where

$$\kappa_\varepsilon = \frac{1}{\max_{z\in\mathbb{R}^d:\|z\|\le 1} z_d}$$

Proof. For the point x_k in the data set, the residual under the assumption $X_d = \boldsymbol{\beta}_0$ is $\varepsilon_k(\beta_0) = \mathrm{D}(x_k, \mathcal{H}_0) = \min_{y\in\mathcal{H}_0}\|x_k - y\|$, where $\mathcal{H}_0 = \{y \in \mathbb{R}^d : y_d = \beta_0\}$ for some $\beta_0 \in \mathbb{R}$. Then, by (4) in Lemma 2.1

$$\varepsilon_k(\beta_0) = \frac{|x_{kd} - \beta_0|}{\|(0,\ldots,0,-1)\|^*}$$

with $\|\cdot\|^*$ the dual norm of $\|\cdot\|$. By definition of the dual norm $\|y\|^* = \max_{z\in\mathbb{R}^d:\|z\|\le 1} z^t y$. Hence, applying such a definition to $y = (0,\ldots,0,-1)$ the result follows. □

From the above result it is easy to see that $\kappa_\varepsilon = 1$, provided that ε_x is induced by any ℓ_τ norm, even for the ℓ_1 and the ℓ_∞ cases. However, as we will see in Section 4, not all the norms have the same κ_ε constant.

Let us introduce the function $f_{\lambda,p}(\beta) := \sum_{i=1}^n \lambda_i\, \varepsilon^p_{(i)}$. Next, with our specifications for Φ, the problem to be solved to obtain Φ^*_0 is:

$$\Phi^*_0 = \kappa_{\boldsymbol{\varepsilon}} \min_{\beta_0\in\mathbb{R}} f_{\lambda,p}(\beta) \tag{17}$$

where $\varepsilon_i = |x_{id} - \beta_0|$ for $i = 1,\ldots,n$.

Solutions to Problem (17) for a given $\beta_0 \in \mathbb{R}$ motivate the introduction of the concept of *ordered median point*. Indeed, β_0 is a $(\lambda,\, p)$-*ordered median point* ($(\lambda,\, p)$-omp in short) if it is an optimal solution to (17).

Some special cases of $(\lambda,\, p)$-omp are well-known and widely used in the so-called Location Analysis literature. If $\lambda_i = 1$ for all $i = 1,\ldots,n$, the $(\lambda,\, 1)$-omp is known to coincide with the median, $\mathrm{median}(x_{1d},\ldots,x_{nd})$, of $\{x_{1d},\ldots,x_{nd}\}$; while the $(\lambda,\, 2)$-omp is the arithmetic mean of the $x_{.d}$-values.

In the general case, i.e., for arbitrary λ and p, the ordered median points do not have closed form expressions (Fernández et al., 2014; 2017), although they have been around in the field of LA for several years (Nickel and Puerto, 1999; 2005). Moreover, they can be obtained, as shown below, to be used in the computation of the goodness of fitting index.

In the following we show how to solve (17) for general choices of non-negative vectors λ and $p \in [1,+\infty)$. Without loss of generality we assume that $x_{1d} \le x_{2d} \le \ldots \le x_{nd}$. Let us denote further by $\alpha_{ik} := \frac{x_{id}+x_{kd}}{2}$ the solution of the equation $\boldsymbol{\varepsilon}^p_i(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^p_k(\boldsymbol{\beta})$ for all $i < k,\ i, k = 1,\ldots,n$ in the range $(x_{1d},\, x_{nd})$. Let $\mathcal{A}$ be the set containing all the $x_{.d}$ and α points and denote by z_k the kth point in $\mathcal{A}$ sorted in non-decreasing sequence. By construction, in the interval $I_k = (z_k, z_{k+1})$ all the functions $\boldsymbol{\varepsilon}^p_i(\boldsymbol{\beta})$ are monotone for all $i = 1,\ldots,n$. Let us denote by $\mathcal{A}_c$ the set of all the critical points of the function $f_{\lambda,\,p}$ in the interval $(x_{1d},\, x_{nd})$ for $p \in (1,+\infty)$.

Theorem 3.2. *For any non-negative vector λ and $p \in (1,\, \infty)$ the set $\mathcal{A} \cup \mathcal{A}_c$ always contains a $(\lambda,\, p)$-omp. For $p = 1$ the set $\mathcal{A}$ always contains a $(\lambda,\, 1)$-omp.*

Proof. For all $\boldsymbol{\beta} \in I_k$, the function $f_{\lambda,\,p}$ for $p \in (1,+\infty)$ is a non-negative linear combination of monotone functions. Therefore, its derivative can vanish in at most one point. This implies that the minimum of $f_{\lambda,\,p}$ is always attained on $\mathcal{A} \cup \mathcal{A}_c$. If $p = 1$ then $f_{\lambda,\,p}$ is a non-negative linear combination of linear functions; and thus the minimum in the interval I_k is attained in one of its extreme points. Hence, the minimum of $f_{\lambda,\,1}$ is attained on $\mathcal{A}$. □

The reader may observe that the implication of the above theorem is that $\hat{\boldsymbol{\beta}}_0$ can be always obtained by a simple enumeration of the set $\mathcal{A} \cup \mathcal{A}_c$ (Observe that the cardinality of this set is $O(n^2)$). Then, $\Phi^*_0 = \kappa_{\boldsymbol{\varepsilon}} \sum_{i=1}^n \lambda_i |x_{id} - \hat{\boldsymbol{\beta}}_0|^p_{(i)}$. Thus, the complexity of computing GoF is essentially the same that the resolution of Problem (1), which must be solved to obtain Φ^*.

Example 3.3. The data considered in this example consists of 47 points in $\mathbb{R}^2$ about stars of the CYG OB1 cluster in the direction of Cygnus (Humphreys, 1978). The first coordinate, X_1, is the logarithm of the effective temperature at the surface of the star and the second one, X_2, is the logarithm of its light intensity. This data set has also been analyzed in Rousseeuw and Leroy (2003) and Yager and Beliakov (2010), among others.

We run the LSS, LAD, LMS and LTS(α) with $\alpha \in \{50,\, 75,\, 90\}$. The obtained lines and the goodness of fitting indices ($\mathrm{GoF}_{\Phi,\,\boldsymbol{\varepsilon}}$) are shown in Fig. 1.

Observe that the LSS and LAD models were not able to adequately fit the data while the others (which are somehow similar) show their better performance against the outliers. Note also that GoF reflects this fact, although it is not clear whether LTS(75) (the one with the largest GoF) is better than the others.

In order to show the behavior of the LTS models and which are the results of their optimal fitting lines, Fig. 2 shows the fitting lines that minimize the 50%, 75% or 90% of the residuals and the points that the corresponding optimization problems discard (filled dots in the subfigures) to reach the fitted lines.

Observe that the percentage of discarded data $(1-\alpha)$ is a key point in LTS models. Several measures are available to determine breakdown points. One of the most widely used measures is the R_α-index (see Atkinson and Cheng, 1999; Hofmann et al., 2010), which is defined as:

$$R_\alpha = \frac{\Phi^*_{LTS(\alpha)}}{\Phi^*_{LSS}} \cdot \frac{n-d}{\lfloor \alpha n \rfloor - d}$$

In Fig. 3, we show the R_α index as a function of α, for the stars dataset. A big slope change in such a function indicates the adequacy of using the corresponding α for the LTS model. As can be observed, R_α has a high-breakdown point in $\alpha = 90\%$ as detected by GoF. Actually, although both indices measure different characteristics of the model (GoF measures the convenience of using the model against the simple constant one and R_α the detection of outliers data in the sample), they have a similar behavior (R_α is similar to $1 - \mathrm{GoF}_{LTS(\alpha)}$). Moreover, the index R_α for the three LTS models can be seen in the table of Fig. 1.

4. Fitting hyperplanes with block-norm residuals

In this section, we present models to compute the parameters of the fitting hyperplanes when distances are assumed to be measured by a block-norm between the points and the closest point in the hyperplane; and the aggregation criterion is considered in the general form given by Problem (1). Recall that a block norm is a norm such that its unit ball is a polytope symmetric with respect to the origin and with non empty interior. Block norms, also referred to as polyhedral norms, play an important role in the measurement of distances in many areas of Operations Research and Applied Mathematics as for instance in Location Analysis or Logistics. They are often used to model real world situations (like measuring highway distances) more accurately than the standard Euclidean norm.

The results in this section will be instrumental to address the general problem of finding hyperplanes with general norms (see Section 5). Using block norms induce linear programming problems and moreover, by its denseness property, any norm can be arbitrarily approximated by block ones (Ward and Wendell, 1985).

We denote by $\|\cdot\|_B$ the norm in $\mathbb{R}^d$ whose unit ball is given by a symmetric with respect to the origin, with non empty inte-

Method	Line	GoF (R_α)
LSS	y=-0.4133 x + 6.7934	0.0442
LAD	y= -0.6931 x + 8.1492	0.0065
LMS	y = 4 x -12.76	0.0765
LTS(50)	y=3.0461 x -8.50	0.3531 (0.1141)
LTS(75)	y= 3.0461 x -8.50	0.4927 (0.2484)
LTS(90)	y = 2.8028x -7.4035	0.4436 (0.4008)

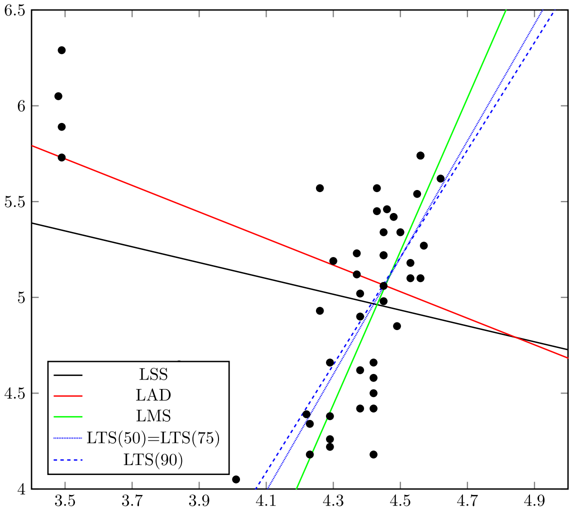


Fig. 1. Optimal lines with the classical methods for the stars data set. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

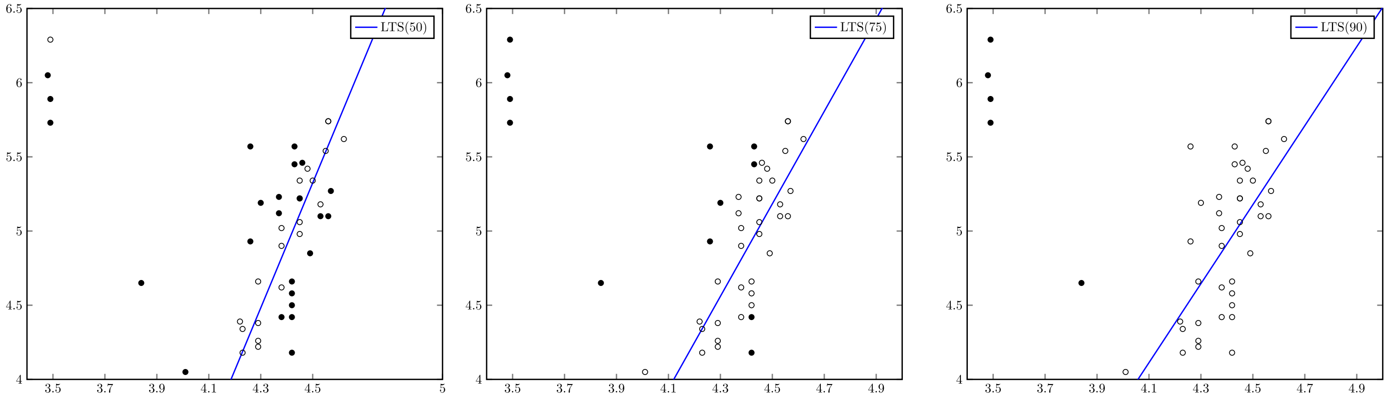


Fig. 2. Estimated models and discarded points (filled dots) in LTS models.

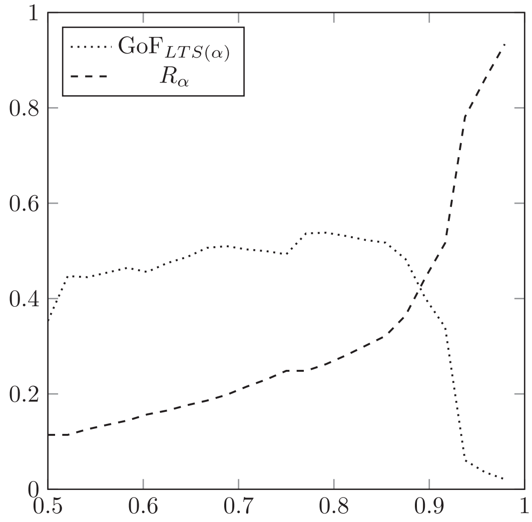


Fig. 3. R_α index for the stars dataset.

rior polytope B, i.e., $B = \{x \in \mathbb{R}^d : \|x\|_B \leq 1\}$. Let $\text{Ext}(B) = \{b_g : g = 1, \ldots, G\}$ be the set of extreme points of B and B^0 the polar set of B which is defined as:

$$B^0 = \{v \in \mathbb{R}^d : v^t b_g \leq 1, g = 1, \ldots, G\}$$

and $\text{Ext}(B^0) = \{b^0_1, \ldots, b^0_{G^0}\}$.

It is well-known (Ward and Wendell, 1980; 1985) that the evaluation of a block norm can be done in terms of the extreme points of the polar set of the polytope B:

$$\|x\|_B = \max\{|x^t b^0_g| : g = 1, \ldots, G^0\}, \quad \text{for all } x \in \mathbb{R}^d. \tag{18}$$

The above expression is a linear program, whose complexity depends on the number of extreme points of B^0. In the case of exponentially many extreme points, one can always resort to column generation techniques to improve the performance of its computation. Special cases of block norms are the Manhattan (ℓ_1) and the Chebyshev (ℓ_∞) norms for adequate choices of the extreme points of the unit balls. Any block norm $\|\cdot\|_B$ in $\mathbb{R}^d$ induces a distance between vectors $x, y \in \mathbb{R}^d$ given by $\text{D}_B(x, y) = \|x - y\|_B$.

Given a set of points $\{x_1, \ldots, x_n\} \subseteq \mathbb{R}^d$ and a polyhedral unit ball B, our goal is to obtain the hyperplane $\mathcal{H}(\boldsymbol{\beta}) = \{y \in \mathbb{R}^d : (1, y^t)\boldsymbol{\beta} = 0\}$ such that the overall distance $\text{D}_B(\cdot, \cdot)$ from the sample to $\mathcal{H}(\boldsymbol{\beta})$ is minimized according to the aggregation function Φ (for $1 \leq p = \frac{r}{s} \in \mathbb{Q}$). That is:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{d+1}} \sum_{i=1}^{n} \lambda_i \boldsymbol{\varepsilon}^p_{(i)}, \tag{RM$_B$}$$

where for any $x \in \mathbb{R}^d$, $\boldsymbol{\varepsilon}_x = \text{D}_B(x, \mathcal{H}(\boldsymbol{\beta}))$, is the "$\|\cdot\|_B$-projection" of x onto the hyperplane $\mathcal{H}(\boldsymbol{\beta})$, and $\boldsymbol{\varepsilon}_{(i)}$ denotes the element in $\{\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n\}$ which is sorted in the ith position (in nondecreasing order).

We recall that according to Eq. (4) in Lemma 2.1, for any polytope B symmetric with respect to the origin and with non empty interior, and $\mathcal{H}(\boldsymbol{\beta}) = \{y^t \in \mathbb{R}^d : (1, y^t)\boldsymbol{\beta} = 0\}$ then $\mathrm{D}_B(x_{-0}, \mathcal{H}(\boldsymbol{\beta})) = \dfrac{|\boldsymbol{\beta}^t x|}{\|\boldsymbol{\beta}_{-0}\|_{B^0}}$, where B^0 is the polar set of B and $x^t = (1, X_1, \ldots, X_d) \in \mathbb{R}^{d+1}$ is a given point.

The following is a simpler valid formulation for the hyperplane location problem with block norm residuals. For a set of linear equations $a_j^t x = b_j$, for $j = 1, \ldots, m$, we denote by $\bigvee_{j=1}^m [a_j^t x = b_j]$ the disjunctive constraint that requires that at least one of the equations $a_j^t x = b_j$ (for $j = 1, \ldots, m$) is satisfied by x.

Theorem 4.1. *Let* $\{x_1, \ldots, x_n\} \subset \mathbb{R}^{d+1}$ *be a set of points and let* $B \subset \mathbb{R}^d$ *be a polytope with* $\mathrm{Ext}(B) = \{b_1, \ldots, b_G\}$. *Then,* ($\mathrm{RM}_B$) *is equivalent to the following disjunctive programming problem*

$$\rho^*(B) := \min \sum_{j=1}^{n} \lambda_j \theta_j \tag{19}$$

$$\text{s.t.} \quad (9)-(13)$$

$$\boldsymbol{\varepsilon}_i \geq \boldsymbol{\beta}^t x_i, \forall i = 1, \ldots, n, \tag{20}$$

$$\boldsymbol{\varepsilon}_i \geq -\boldsymbol{\beta}^t x_i, \forall i = 1, \ldots, n, \tag{21}$$

$$\boldsymbol{\beta}_{-0}^t b_g \leq 1, \ \forall g = 1, \ldots, G, \tag{22}$$

$$\bigvee_{g=1}^{G} \left[\boldsymbol{\beta}_{-0}^t b_g = 1\right], \tag{23}$$

$$\gamma_{ik} \in \{0, 1\}, \omega_{ik} \geq 0, \ \Delta_k < 0,$$
$$z_{ik}, \ t_k \geq 0, \quad i, k = 1, \ldots, n, \ \Delta_k > 0$$
$$\boldsymbol{\beta} \in \mathbb{R}^{d+1}, \ \varepsilon_i \geq 0, \ i = 1, \ldots, n.$$

Proof. Let us denote by $\boldsymbol{\varepsilon}_i = \mathrm{D}_B(x_i, \mathcal{H}(\boldsymbol{\beta}))$. By Lemma 2.1, $\boldsymbol{\varepsilon}_i = \frac{|\boldsymbol{\beta}^t x_i|}{\|\boldsymbol{\beta}_{-0}\|_{B^0}}$. Let $\boldsymbol{\beta}^* \in \mathbb{R}^{d+1}$ be an optimal solution of (RM_B) with $\boldsymbol{\beta}^*_{-0} \neq 0$. Then, $\boldsymbol{\beta}' = \frac{\boldsymbol{\beta}^*}{\|\boldsymbol{\beta}_{-0}\|_{B^0}}$ is also an optimal solution of (RM_B) with $\|\boldsymbol{\beta}'_{-0}\|_{B^0} = 1$. Thus, there is an optimal solution of (RM_B), $\boldsymbol{\beta}$, that verifies $\mathrm{D}_B(x_{-0}, \mathcal{H}(\boldsymbol{\beta})) = |\boldsymbol{\beta}^t x|$ for any $x^t = (1, x_1, \ldots, x_d) \in \mathbb{R}^{d+1}$. Therefore, we can assume that $\|\boldsymbol{\beta}_{-0}\|_{B^0} = 1$, hence $\boldsymbol{\varepsilon}_i = |\boldsymbol{\beta}^t x_i|$ (constraints (20) and (21)). Since $(B^0)^0 = B$ then $\|\boldsymbol{\beta}_{-0}\|_{B^0} = \max\{|\sum_{i=1}^d \beta_i b_{gi}| : g = 1, \ldots, G\}$. Hence, there exists $g_0 \in \{1, \ldots, G\}$ such that $\|\boldsymbol{\beta}_{-0}\|_{B^0} = 1$ (disjunctive constraint (23)) and thus $\sum_{k=1}^d \beta_k b_{gk} \leq \sum_{k=1}^d \beta_k b_{g_0 k} = 1$ (constraint (22)). (Note that absolute values do not need to be taken explicitly into account since if $b_g \in \mathrm{Ext}(B)$, then $-b_g \in \mathrm{Ext}(B)$.) □

The above problem can be equivalently written as a Mixed Integer Second Order Cone Optimization (MISOCO) problem once constraints (9) are transformed, using the result in Remark 2.5, and binary variables are added to decide which g_0 is chosen to verify constraint (23). By the same token, this problem can be also equivalently rewritten as G (recall that G is the cardinality of $\mathrm{Ext}(b)$) different Second Order Cone Programming Problems (SOCP) (each of them fixed to verify one of the disjunctive constraints). Furthermore, mixed integer non linear disjunctive programming techniques (see, e.g., Balas, 1979, Lee and Grossmann, 2000) may be used to solve the corresponding problem. Based in the above discussion, the following is another valid MINLP formulation for (RM_B).

Corollary 4.2. *Let* $\{x_1, \ldots, x_n\} \subset \mathbb{R}^{d+1}$ *be a set of points and let* $B \subset \mathbb{R}^d$ *be a polytope with* $\mathrm{Ext}(B) = \{b_1, \ldots, b_G\}$. *Then,* (19) *is equivalent to the following problem:*

$$\rho^*(B) := \min \sum_{j=1}^{n} \lambda_j \theta_j \tag{24}$$

$$\text{s.t.} \quad (9)-(13)$$

$$\boldsymbol{\varepsilon}_i \geq \boldsymbol{\beta}_h^t x_i, \forall i = 1, \ldots, n, h = 1, \ldots, G, \tag{25}$$

$$\boldsymbol{\varepsilon}_i \geq -\boldsymbol{\beta}_h^t x_i, \forall i = 1, \ldots, n, h = 1, \ldots, G, \tag{26}$$

$$\boldsymbol{\beta}_{-0h}^t b_g \leq 1, \ \forall g = 1, \ldots, G, h = 1, \ldots, G, \tag{27}$$

$$\boldsymbol{\beta}_{-0h}^t b_h = \xi_h, h = 1, \ldots, G, \tag{28}$$

$$\sum_{h=1}^{G} \xi_h = 1, \tag{29}$$

$$\boldsymbol{\beta}_h \in \mathbb{R}^{d+1}, \xi_h \in \{0, 1\}, \forall h = 1, \ldots, G,$$
$$\gamma_{ik} \in \{0, 1\}, \omega_{ik} \geq 0, \ \Delta_k < 0,$$
$$z_{ik}, \ t_k \geq 0, \quad i, k = 1, \ldots, n, \ \Delta_k > 0$$
$$\varepsilon_i \geq 0, \ i = 1, \ldots, n.$$

Some special cases for the aggregation function Φ allow us even simpler formulations reducing considerably the computational complexity of the problems. In particular, when $\lambda_i = 1$ for all $i = 1, \ldots, n$, the integer variables representing ordering (w_{ij}) can be removed from the above formulation.

The following result permits to consider polyhedral norms which are *dilations* of other polyhedral norms, i.e., polyhedral norms $\|\cdot\|_{\mu B}$ for some bounded polyhedron B and $\mu > 0$ ($\mu B = \{\mu \ \ z : z \in B\}$). It will be very useful in the next section when we approximate the problem of locating hyperplanes with general norms by problems with polyhedral ones.

Lemma 4.3. *Let* $\overline{B}$ *be a polytope and* $\mu > 0$. *Then, if* $\boldsymbol{\beta}^*$ *is an optimal solution for Problem* (24) *for* $B = \overline{B}$, $\widehat{\boldsymbol{\beta}} = \frac{1}{\mu}\boldsymbol{\beta}^*$ *is an optimal solution for* (24) *when* $B = \mu\overline{B}$. *Moreover,* $\rho^*(\mu\overline{B}) = \frac{1}{\mu^p}\rho^*(\overline{B})$.

Proof. It is sufficient to observe that for any $\boldsymbol{\beta} \in \mathbb{R}^{d+1}$:

$$\begin{aligned} &\|(\beta_1, \ldots, \beta_d)\|_{\mu\overline{B}^0} \\ &\quad = \max\{|\mu b_g^t \boldsymbol{\beta}^t| : g = 1, \ldots, G\} \\ &\quad = \mu \max\{|b_g^t \boldsymbol{\beta}^t| : g = 1, \ldots, G\} = \mu\|(\beta_1, \ldots, \beta_d)\|_{\overline{B}^0}. \end{aligned}$$

Since $\Phi_{\mu\overline{B}}(\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n) = \frac{1}{\mu^p}\Phi_{\overline{B}}(\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)$, we get the relation between the optimal values. Let $\boldsymbol{\beta}^*$ be an optimal solution of (24). Then, $\frac{1}{\mu}\boldsymbol{\beta}^*$ is clearly a feasible solution to Problem (24) when $B = \mu\overline{B}$ since $\|(\frac{1}{\mu}\beta_1^*, \ldots, \frac{1}{\mu}\beta_d^*)\|_{\mu\overline{B}^0} = \|(\beta_1^*, \ldots, \beta_d^*)\|_{\overline{B}^0} = 1$. □

In order to compute GoF for solutions to problems with block-norm residuals, note that the one dimensional Problem (16) does depend on Φ and also on the residuals through κ_ε. Let us denote by κ_B the constant κ_ε when the residuals ε_x are defined as the block-norm projection with unit ball given by the polytope B.

Corollary 4.4. *Let* $B \subset \mathbb{R}^d$ *be a polytope. The Goodness of Fitting index,* GoF, *when the residuals are defined as the block-norm distance with unit ball* B, *can be computed as:*

$$\mathrm{GoF}_{\Phi,\varepsilon} = 1 - \frac{\Phi^*}{\sum_{i=1}^n |x_{id} - ((\lambda, p) - \mathrm{omp}(x_{\cdot d}))|^p} \cdot \max_{g=1,\ldots,G} |b_{gd}|,$$

Method (Φ, ε)	Optimal Line	$\mathrm{GoF}_{\Phi,\varepsilon}$
(SUM, ℓ_1)	$y = 7x - 25.81$	0.6505853
(SUM, ℓ_∞)	$y = 5.25x + -18.1425$	0.7009688
(SUM, Hex)	$y = 7x - 25.81$	0.6505853
(MAX, ℓ_1)	$y = -3.230769x + 18.77577$	0.5336373
(MAX, ℓ_∞)	$y = -3.230769x + 18.77577$	0.6438685
(MAX, Hex)	$y = -3.230769x + 18.77577$	0.6438685
(kC, ℓ_1)	$y = -4.307692x + 23.03346$	0.4628481
(kC, ℓ_∞)	$y = -2.493333x + 15.67113$	0.5921635
(kC, Hex)	$y = 7.642857x + -28.67929$	0.8317972
(AkC, ℓ_1)	$y = 5.6x - 19.804$	0.8443055
(AkC, ℓ_∞)	$y = 4.869565x - 16.41565$	0.8426523
(AkC, Hex)	$y = 5.473684x - 19.28316$	0.6431602

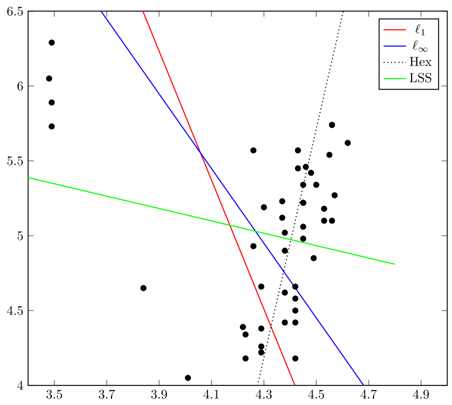


Fig. 4. Optimal lines obtained with block-norm residuals for the stars data set. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

where (λ, p)*-omp*$(x_{\cdot d})$ *is the solution to the Problem (16) with residuals measured with the polyhedral norm with unit ball B.*

Proof. By Lemma 3.1 the goodness of fitting index $\mathrm{GoF}_{\Phi,\,\varepsilon}$ can be computed as:

$$\mathrm{GoF}_{\Phi,\varepsilon} = 1 - \frac{\Phi^*}{\min_{\beta_0\in\mathbb{R}} \Phi(\kappa_B|x_{1d} - \beta_0|, \ldots, \kappa_B|x_{nd} - \beta_0|)}, \qquad (30)$$

where $\kappa_B = \frac{1}{\max_{z\in B} z_d}$.

Observe that since B is a polytope then the above maximum is attained in an extreme point of B and thus $\kappa_B = \frac{1}{\max_{g=1,\ldots,G} b_{gd}}$.

Next, Problem (16) in this case can be expressed as:

$$\kappa_B \cdot \min_{\beta_0\in\mathbb{R}} \sum_{i=1}^{n} \lambda_i |x_{\cdot d} - \beta_0|^p_{(i)}.$$

Recall that this is a (λ, p) *Ordered median problem* and that its optimal solution, a (λ, p)-omp, can be easily obtained by the result in Theorem 3.2. Replacing the optimal solution to this problem in (30) it results in:

$$\mathrm{GoF}_{\Phi,\varepsilon} = 1 - \frac{\Phi^*}{\sum_{i=1}^n |x_{id} - ((\lambda, p) - \mathrm{omp}(x_{\cdot d}))|^p} \cdot \max_{g=1,\ldots,G} |b_{gd}|. \quad \square$$

Note that for $\lambda = (1, \ldots, 1)$ the $(\lambda,\ 1)$-omp is the standard median point and thus the expression $\sum_{i=1}^n |x_{id} - \mathrm{median}(x_{\cdot d})|$ is what it is usually called the *mean absolute deviation with respect to the median*.

The same dataset used in Example 3.3 allows us to show the expressions of the optimal fitting hyperplanes when different block-norm residuals are considered:

Example 4.5. We consider again the stars data used in Example 3.3. In this case, we run our implementation in R for ℓ_1-norm, ℓ_∞-norm and hexagonal norm (as the one used in Nickel and Puerto (2005) with $\mathrm{Ext}(B) = \{\pm(2, 0), \pm(2, 2), \pm(-1, 2)\}$) residuals. This last choice is included only for illustrative purposes of the presented methodology and by its applicability in LA, although its statistical meaning may need further investigation. We also note in passing that the use of different metrics, based on geodesic of the considered space, is natural in geodesic regression (Fletcher, 2013). We use four different criteria: overall SUM ($\lambda = (1, \ldots, 1)$ and $p = 1$), MAXimum ($\lambda = (1, 0, \ldots, 0)$ and $p = 1$), K-centrum ($\lambda = (\overbrace{0, \ldots, 0}^{K}, \overbrace{1, \ldots, 1}^{n-K})$) for $K = \lfloor 0.75n \rfloor$ (the model will minimize the sum of the 25% greatest residuals) and anti-K-centrum ($\lambda = (\overbrace{1, \ldots, 1}^{K}, \overbrace{0, \ldots, 0}^{n-K})$) for $K = \lfloor 0.5n \rfloor$ (the model will minimize the sum of the 50% smallest residuals). The results for all the combinations and the graph for the K-centrum lines are shown in Fig. 4.

Note that different situations may happen when running the different models: in the case of the SUM criterion the models for ℓ_1 and hexagonal residuals coincide; for the MAX criterion the three optimal lines are the same, and for the K-centrum and anti-K-centrum the three models are different. Furthermore, even in the case when the models coincide, one may have different goodness of fitting indices due to the different way of measuring distances (see the ℓ_1 and hexagonal residuals for the MAX criterion).

From the above, we observed that the GoF are not comparable when different residuals are used in the models since the value given to the residuals (both with respect to the best model and with respect to the simplified model with only intercept) is different. Thus, the generalized coefficient allows us to compare the goodness of fitting between models provided that the distance (to measure the residuals) and the aggregation criterion are fixed.

5. Fitting hyperplanes with ℓ_τ distances

In this section we deal with the general problem of locating a hyperplane with respect to a set of points and we present a suitable mathematical programming formulation for computing the optimal hyperplanes when the residuals are defined as ℓ_τ, $\tau \geq 1$, distances. Recall that for any $z = (z_1, \ldots, z_d)^t \in \mathbb{R}^d$ the ℓ_τ-norm, $\tau \geq 1$, is defined as:

$$\|z\|_\tau = \begin{cases} \left(\displaystyle\sum_{k=1}^{d} |z_k|^\tau\right)^{\frac{1}{\tau}} & \text{if } \tau < \infty, \\ \displaystyle\max_{k=1,\ldots,d} \{|z_k|\} & \text{if } \tau = \infty. \end{cases}$$

From this norm we denote by $\mathrm{D}_{\ell_\tau}(z, y) = \|z - y\|_\tau$ the ℓ_τ-distance between the points $z, y \in \mathbb{R}^d$. The well-known Euclidean distance, that measures the straight line distance between points, is the ℓ_2-norm in this family. Note that the extreme cases of ℓ_1 and ℓ_∞ represent both block and ℓ_τ-norms, since their unit balls are polytopes but also fit within the family of ℓ_τ-norms.

We recall that according to Eq. (4) in Lemma 2.1, for any $\tau = \frac{r}{s} \in \mathbb{Q}$ with $r \geq s \in \mathbb{Z}_+$, $\gcd(r, s) = 1$ and $\mathcal{H}(\boldsymbol{\beta}) = \{y^t \in \mathbb{R}^d : (1, y^t)\boldsymbol{\beta} = 0\}$, then $\mathrm{D}_\tau(z, \mathcal{H}(\boldsymbol{\beta})) = \frac{|\boldsymbol{\beta}^t z|}{\|\boldsymbol{\beta}_{-0}\|_\nu}$, where ν is such that $\frac{1}{\tau} + \frac{1}{\nu} = 1$ (for $\tau = 1$, $\nu = \infty$ while for $\tau = \infty$, $\nu = 1$).

In this section we assume that the residuals are defined as the shortest distance from the points to the fitted hyperplane, namely, for a given point $\hat{x} = (1, \hat{x}_1, \ldots, \hat{x}_d)^t$ the residual is: $\varepsilon_{\hat{x}}(\boldsymbol{\beta}) = \mathrm{D}_\tau(\hat{x}_{-0}, \mathcal{H}(\boldsymbol{\beta}))$.

Let $\{x_1, \dots, x_n\} \subset \mathbb{R}^{d+1}$ be a given set of points, $\lambda \in \mathbb{R}^n$, $\tau = \frac{r}{s} \in \mathbb{Q}$ with $r > s \in \mathbb{N}$ and $\gcd(r, s) = 1$, and $\|\cdot\|_\tau$, a ℓ_τ-norm in $\mathbb{R}^d$. It follows from the discussion above that under these hypotheses, Problem (1) is equivalent to the following mathematical programming problem:

$$\Phi^*_{\ell_\tau} := \min \sum_{j=1}^{n} \lambda_j \theta_j \tag{31}$$

$$\begin{aligned} \text{s.t.} \quad & (8)-(13), (20)-(21), \\ & \|\boldsymbol{\beta}_{-0}\|_\nu = 1, \qquad (32) \\ & \gamma_{ik} \in \{0, 1\}, \omega_{ik} \geq 0, \ \Delta_k < 0, \\ & z_{ik}, \ t_k \geq 0, \quad i, k = 1, \dots, n, \ \Delta_k > 0 \\ & \boldsymbol{\beta} \in \mathbb{R}^{d+1}, \ \varepsilon_i \geq 0, \ i = 1, \dots, n. \end{aligned}$$

Note that the above problem is nonconvex for $1 < \tau < \infty$ because of the binary variables and constraint (32). One could try to solve Problem (31) using algorithms available in different nonlinear optimization solvers, although no guarantee of optimality is provided (e.g., NLOPT, BARON, Minotaur, ...). In what follows we describe an accurate approximation alternative based on the results in Section 4.

Let P be a polyhedron such that $P \subset \mathcal{B} = \{z \in \mathbb{R}^d : \|z\|_\nu \leq 1\}$, and denote by $r_P = \sup_{\|z\|_P=1} \|z\|_\nu$ (note that by construction $r_P \leq 1$). Observe that r_P is the radius of the smallest ℓ_ν-ball containing P. In addition, let Q be a polyhedron such that $\mathcal{B} \subset Q$, and denote by $R_Q = \inf_{\|z\|_Q=1} \|z\|_\nu$ (note that by construction $R_Q \geq 1$). In this case R_Q is the radius of the largest ℓ_ν-ball contained in Q.

For a generic polyhedron P, let $\boldsymbol{\varepsilon}_P = (\varepsilon_{1,P}, \dots, \varepsilon_{n,P})^t$, with $\varepsilon_{i,P} = D_P(x_{i,-0}, \mathcal{H})$, $i = 1, \dots, n$. Analogously, let $\boldsymbol{\varepsilon}_{\ell_\tau} = (\varepsilon_{1,\ell_\tau}, \dots, \varepsilon_{n,\ell_\tau})^t$, with $\varepsilon_{i,\ell_\tau} = D_{\ell_\tau}(x_{i,-0}, \mathcal{H})$, $i = 1, \dots, n$. Let $\delta = \frac{r}{s} \in \mathbb{Q}$ with $r, s \in \mathbb{Z} \backslash \{0\}$ with $gcd(r, s) = 1$.

The following result states the relationship between the objective values obtained when using either ℓ_τ or the block-norms induced by P and Q to define the residuals in our models.

Theorem 5.1. *Let* $\lambda_1, \dots, \lambda_n \geq 0$ *and the aggregation function* $\Phi(\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_n) = \sum_{i=1}^n \lambda_i \boldsymbol{\varepsilon}^\delta_{(i)}$ *then:*

$$\Phi(\boldsymbol{\varepsilon}_P) \leq \Phi(\boldsymbol{\varepsilon}_{\ell_\tau}) \leq \frac{1}{r_P^\delta} \Phi(\boldsymbol{\varepsilon}_P) \tag{33}$$

$$\frac{1}{R_Q^\delta} \Phi(\boldsymbol{\varepsilon}_Q) \leq \Phi(\boldsymbol{\varepsilon}_{\ell_\tau}) \leq \Phi(\boldsymbol{\varepsilon}_Q) \tag{34}$$

Proof. By the relations between the norms, it is clear that $\|z\|_P \geq \|z\|_\nu \geq r_P \|z\|_P$. Let $\mathcal{H}(\boldsymbol{\beta}) = \{z \in \mathbb{R}^d : (1, z^t)\boldsymbol{\beta} = 0\}$. Then, for any $x \in \mathbb{R}^d$, the above relationships imply the following inequalities relating the distances with respect to $\| \cdot \|_{P0}$-residuals and $\|\cdot\|_\tau$-residuals:

$$\mathrm{D}_{P0}(x_{-0}, \mathcal{H}(\boldsymbol{\beta})) = \frac{|\boldsymbol{\beta}^t x|}{\|\boldsymbol{\beta}_{-0}\|_P} \leq \frac{|\boldsymbol{\beta}^t x|}{\|\boldsymbol{\beta}_{-0}\|_\nu} \leq \mathrm{d}_\tau(x_{-0}, \mathcal{H}(\boldsymbol{\beta}))$$

and

$$\mathrm{D}_\tau(x_{-0}, \mathcal{H}(\boldsymbol{\beta})) = \frac{|\boldsymbol{\beta}^t x|}{\|\boldsymbol{\beta}_{-0}\|_\nu} \leq \frac{|\boldsymbol{\beta}^t x|}{r_P \|\boldsymbol{\beta}_{-0}\|_P} \leq \frac{1}{r_P} \mathrm{D}_{P0}(x_{-0}, \mathcal{H}(\boldsymbol{\beta}))$$

Let us consider the aggregation criterion $\Phi(\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_n) = \sum_{i=1}^n \lambda_i \boldsymbol{\varepsilon}^\delta_{(i)}$. Its evaluation with respect to the residuals computed with the polyhedral norm with unit ball P and the ℓ_τ-norm, namely $\varepsilon_{i,P} = \mathrm{D}_P(x_{i,-0}, \mathcal{H}(\boldsymbol{\beta}))$ and $\varepsilon_{i,\ell_\tau} = \mathrm{D}_\tau(x_{i,-0}, \mathcal{H}(\boldsymbol{\beta}))$ for all $i = 1, \dots, n$, satisfies:

$$\Phi(\boldsymbol{\varepsilon}_P) \leq \Phi(\boldsymbol{\varepsilon}_{\ell_\tau}) \leq \frac{1}{r_P^\delta} \Phi(\boldsymbol{\varepsilon}_P).$$

This equation proves (33).

Next, by definition of Q, it is clear that $\|z\|_Q \leq \|z\|_\nu \leq R_Q \|z\|_Q$. Now, using an argument similar to the one above we conclude that

$$\begin{aligned} \mathrm{D}_Q(x_{-0}, \mathcal{H}(\boldsymbol{\beta})) = \frac{|\boldsymbol{\beta}^t x|}{\|\boldsymbol{\beta}_{-0}\|_Q} &\geq \frac{|\boldsymbol{\beta}^t x|}{\|\boldsymbol{\beta}_{-0}\|_\nu} \geq \mathrm{D}_\tau(x_{-0}, \mathcal{H}(\boldsymbol{\beta})) \\ &= \frac{|\boldsymbol{\beta}^t x|}{\|\boldsymbol{\beta}_{-0}\|_\nu} \geq \frac{|\boldsymbol{\beta}^t x|}{R_Q \|\boldsymbol{\beta}_{-0}\|_\nu} \geq \frac{1}{R_Q} \mathrm{D}_Q(x_{-0}, \mathcal{H}(\boldsymbol{\beta})). \end{aligned}$$

From these inequalities it clearly follows (34). □

Let P_N be a symmetric with respect to the origin polytope with N vertices, $\{p_1, \dots, p_N\}$, inscribed in the ℓ_ν hypersphere $\mathcal{B} = \{z \in \mathbb{R}^d : \|z\|_\nu = 1\}$ and let r_{P_N} be the radius of the smallest ℓ_ν ball centered at the origin containing P_N. Let $R_{Q_N} = \frac{1}{r_{P_N}}$ and denote by Q_N the R_{Q_N}-dilation of P_N. By construction $P_N \subset \mathcal{B} \subset Q_N$. Hence, for the globalizing function $\Phi(\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_n) = \sum_{i=1}^n \lambda_i \boldsymbol{\varepsilon}^\delta_{(i)}$, by Theorem 5.1, we get that:

$$\begin{aligned} &\max \left\{ \Phi(\boldsymbol{\varepsilon}_{P_N}), \frac{1}{R_{Q_N}^\delta} \Phi(\boldsymbol{\varepsilon}_{Q_N})) \right\} \\ &\quad \leq \Phi(\boldsymbol{\varepsilon}_{\ell_\tau}) \leq \min \left\{ \Phi(\boldsymbol{\varepsilon}_{Q_N}), \frac{1}{r_{P_N}^\delta} \Phi(\boldsymbol{\varepsilon}_{P_N}) \right\} \end{aligned}$$

Furthermore, by Lemma 4.3, since Q_N is a dilation of P_N, both problems have the same optimal solutions and $\Phi(\boldsymbol{\varepsilon}_{P_N}) = r_P^\delta \Phi(\boldsymbol{\varepsilon}_{Q_N})$. Hence,

$$\Phi(\boldsymbol{\varepsilon}_{P_N}) \leq \Phi(\boldsymbol{\varepsilon}_{\ell_\tau}) \leq \frac{1}{r_{P_N}^\delta} \Phi(\boldsymbol{\varepsilon}_{P_N}).$$

It is clear from its definition that r_{P_N} determines the approximation error whenever a ℓ_ν-norm is replaced by a polyhedral norm with unit ball P_N and it can be explicitly computed.

Lemma 5.2. *Let* $P = \{z \in \mathbb{R}^d : a_i x \leq b_i, i = 1, \dots, N\}$ *be a polytope, then:*

$$r_P = \max_{i=1,\dots,N} \frac{b_i}{\|a_i\|_\tau}.$$

Proof. First, note that $r_P = \sup_{\|z\|_P=1} \|z\|_\nu = \max_{\|z\|_P=1} \|z\|_\nu$ by the compactness of P. Thus, r_P is the ℓ_ν-inradius of P. Next, by Mangasarian (1999), the radius of a ℓ_ν ball centered at the origin and reaching the facet $\{x \in \mathbb{R}^d : a_i^t x \leq b\}$ of P is the ℓ_ν projection of the origin onto that facet, namely $\frac{|b_i|}{\|a_i\|_\tau}$. Hence, r_P is the maximum of those distances among the N facets defining P. □

Next, we can obtain from the above discussion a lower bound for $\Phi^*_{\ell_\tau}$, the optimal value of Problem (31). Indeed, it follows that

$$\rho^* \leq \Phi^*_{\ell_\tau} \leq \frac{1}{r_P^p} \rho^*, \tag{35}$$

where

$$\rho^* := \min \sum_{j=1}^{n} \lambda_j \theta_j \tag{36}$$

$$\begin{aligned} \text{s.t. } & (8)-(13) \\ & \boldsymbol{\varepsilon}_i \geq |\boldsymbol{\beta}^t x_i|, \qquad \forall i = 1, \dots, n, \qquad (37) \\ & \|\boldsymbol{\beta}_{-0}\|_{P_N} = 1, \qquad (38) \\ & \gamma_{ik} \in \{0, 1\}, \omega_{ik} \geq 0, \ \Delta_k < 0, \\ & z_{ik}, \ t_k \geq 0, \quad i, k = 1, \dots, n, \ \Delta_k > 0 \\ & \boldsymbol{\beta} \in \mathbb{R}^{d+1}, \ \varepsilon_i \geq 0, \ i = 1, \dots, n. \end{aligned}$$

Table 1
Estimated models with minisum criterion in Example 3.3.

τ	N	$\widehat{\boldsymbol{\beta}}$	Φ^*	GoF	R_P	r_P	Time	SD
1.5	16	(36.87, −1, 0.14)	77.1857	0.6505	0.9848	1.015	1.0	7.26×10^{-5}
	80	(36.84, −0.99, 0.14)	77.1324	0.6508263	0.9993	1.0006	1.97	6.06×10^{-6}
	320	(36.83, −0.99, 0.14)	77.1117	0.6509203	0.9999	1.0000	14.16	9.41×10^{-9}
2	16	(36.87, −1, 0.14)	77.1857	0.6505	0.9807	1.0195	1.04	7.87×10^{-3}
	80	(36.19, −0.98, 0.14)	76.3703	0.654276	0.9922	1.0007	2.01	1.91×10^{-7}
	320	(36.19, −0.98, 0.14)	76.3700	0.654277	0.9999	1.0000	16.53	1.64×10^{-7}
3	16	(34.35, −0.96, 0.16)	74.7283	0.6617	0.9801	1.0202	1.07	4.56×10^{-3}
	80	(34.09, −0.95, 0.16)	74.1627	0.66427	0.9992	1.0007	2.04	3.50×10^{-6}
	320	(34.08, −0.95, 0.16)	74.1468	0.6643	0.9999	1.0000	17.48	4.68×10^{-10}

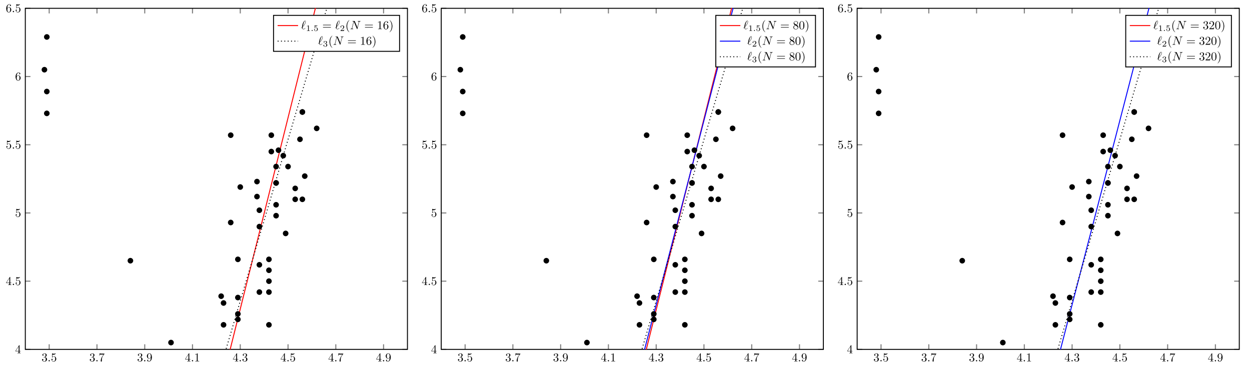


Fig. 5. Estimated lines for the data in Example 3.3 approximating by a {16, 80, 320}-gon. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

For a given finite set of input points, the proposed polyhedral approximation of a ℓ_τ-norm may be exact for an adequate choice of the block-norm. Indeed, this norm must have as fundamental directions the vectors defining the optimal ℓ_τ-projections of each input point onto the optimal hyperplane.

Corollary 5.3. *For any data set* $\{x_1, \ldots, x_n\} \subset \mathbb{R}^{d+1}$ *and any* ℓ_τ*-norm with* $1 < \tau < +\infty$ *there exists a polyhedral norm* $\|\cdot\|_B$ *whose unit ball B has at most* $2n$ *extreme points and such that the optimal values of the problems (31) and (24) coincide.*

In Love and Morris (1972) the authors propose a measure of the quality of the approximation of a given norm by another norm. This measure was defined in order to quantify the approximation errors when modeling road distances between cities. We adapt this measure to evaluate the approximation errors induced whenever the ℓ_τ-norm is replaced by the polyhedral norm with unit ball the polytope P:

$$\mathrm{SD}_{\tau,P}(\boldsymbol{\beta}; \{x_1, \ldots, x_n\}) = \sum_{\substack{i=1 \\ \mathrm{D}_\tau(x_i, \boldsymbol{\beta})>0}}^{n} \frac{(\mathrm{D}_\tau(x_i, \boldsymbol{\beta}) - \mathrm{D}_P(x_i, \boldsymbol{\beta}))^2}{\mathrm{D}_\tau(x_i, \boldsymbol{\beta})}$$

Example 5.4. Let us consider again the stars data from Example 3.3. We run now the models using as aggregation criterion the overall sum of the residuals ($\Phi = SUM$) and the errors are the ℓ_τ projections of the points onto the optimal line, for $\tau \in \{1.5,\ 2,\ 3\}$. The estimations for the aggregation criterion $\Phi = SUM$ and their goodness of fitting ($\mathrm{GoF}_{\Phi,\ \boldsymbol{\varepsilon}}$) are shown in Table 1. The lines are drawn in Fig. 5.

Observe that for this data set, getting high accuracy for the ℓ_τ-norm residual problems is possible using small number of vertices (N) in the approximation by polyhedral norms. As expected, increasing the number of vertices improves the accuracy, at the price of increasing the computation times.

We also computed the optimal lines for different aggregation criteria ($\Phi \in \{\mathrm{SUM}, \mathrm{MAX}, \mathrm{kC}, \mathrm{AkC}\}$) with ℓ_τ residuals, $\tau \in \{1.5,\ 2,\ 3\}$, using the polyhedral approximation approach with $N = 480$ vertices. The results are shown in Table 2. The reader may observe from these results that the approximation error, although tiny, depends both of the chosen residuals and aggregation criteria.

Finally, we compare our approximation scheme for ℓ_τ residuals, on this data set, with other available implementations. Orthogonal Distance Regression (ODR) is a particular case of our general framework where ℓ_2 residuals are chosen and Φ is the sum of squares criterion (note that both approaches coincide when the coefficient of the dependent coordinate is non zero while such an assumption is not imposed in our models). The package `pracma` in R permits to compute ODR by using an approximated iterative procedure (see Boggs and Rogers, 1990). The models obtained with both approaches are shown in the following table. We observe that, for this data set, our approach to approximate ℓ_τ distances by polyhedral norms (with $N = 320$ vertices) has a better performance on the global error measure of the models (although the models obtained by both methods are almost the same):

	ODR	SOS-ℓ_2 (SD=9.93×10^{-11})
Model	$y = -7.05736x + 35.42935$	$y = -7.098062x + 35.60477$
Global Residuals	3.959383	3.662783

6. Experiments

In this section we report the computational results of the proposed methodology. We combine several aggregation criteria and norm-based residuals to find different optimal hyperplanes. Our aim is to show the powerfulness of modern mathematical programming in its application to the considered problem and to compare the behavior of different models rather than gaining insights into their statistical meaning, which is beyond the scope of this paper. Our formulations have been coded in Gurobi 6.0 under R and executed in a PC with an Intel Core i7 processor at 2 × 2.40 GHz and 4 GB of RAM. Overall, we compared 42 methods which results from: **1)** the combination of 7 aggregation criteria: SUM (summation), MAX (maximum), MED (median), kC (summation of the k

Table 2
Optimal lines for different criteria and ℓ_τ residuals of Example 5.4.

		$\ell_{1.5}$	ℓ_2	ℓ_3
SUM	Line	$y = 5.92x - 21.1016$	$y = 6.75x - 24.6975$	$y = 7x - 25.81$
	GoF	0.6643	0.6542	0.6509
	SD	3.36×10^{-10}	1.73×10^{-10}	1.65×10^{-9}
MAX	Model	$y = -3.2307x + 18.7757$	$y = -3.2307x + 18.7757$	$y = -3.2307x + 18.7757$
	GoF	0.5805	0.5544	0.5381
	SD	4.07×10^{-14}	1.90×10^{-12}	3.85×10^{-13}
kC	Model	$y = -2.8133x + 16.9367$	$y = -3.1756x + 18.5100$	$y = -4.3076x + 23.0334$
	GoF	0.5111	0.4790	0.4650
	SD	3.51×10^{-13}	7.53×10^{-10}	9.70×10^{-10}
AkC	Model	$y = 6.75x - 25.0875$	$y = 6.5555x - 24.1533$	$y = 5.175x - 17.7146$
	GoF	0.8092	0.82512	0.8217
	SD	7.15×10^{-10}	2.10×10^{-9}	5.49×10^{-10}

Table 3
Combinations of chosen aggregation criteria and residuals.

Aggregation criteria		Residuals
SUM	$\sum_{i=1}^{n} \boldsymbol{\varepsilon}_i$	V
MAX	$\max_{i=1,\ldots,n} \boldsymbol{\varepsilon}_i$	ℓ_1
MED	$\text{median}(\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)$	ℓ_∞
kC	$\sum_{i=1}^{\lfloor 0.5n \rfloor} \boldsymbol{\varepsilon}_{(i)}$	$\ell_{\frac{3}{2}}$
AkC	$\sum_{i=\lfloor 0.5n \rfloor+1}^{n} \boldsymbol{\varepsilon}_{(i)}$	ℓ_2
SOS	$\sum_{i=1}^{n} \boldsymbol{\varepsilon}_i^2$	ℓ_3
1.5SUM	$\sum_{i=1}^{n} \boldsymbol{\varepsilon}_i^{\frac{3}{2}}$	

largest), AkC (summation of the k smallest), SOS (sum of squares) and 1.5SUM (sum of residuals raised to the power of $\frac{3}{2}$); and **2)** six different modes to measure the residuals: V (vertical distance) and ℓ_τ (ℓ_τ-norm distance for $\tau = 1, \frac{3}{2}, 2, 3, +\infty$). See Table 3.

All experiments were run with a CPU time limit of one hour. The necessary computing times depend very much of the chosen model and, for our instances, range from a few seconds, for the simplest ones, to close to one hour, for the most difficult ones.

We tested the models on two different types of datasets: randomly generated data and a real-word benchmark dataset. The first one will allows us to analyze the performance of the different models in terms of their ability to detect the trend of the dataset. The second one permits to check whether the use of different aggregation criteria and residuals is useful in practice.

6.1. Synthetic experiments

The first set of results is built on randomly generated points following a similar scheme to those proposed in Bertsimas and Mazumder (2014). We generated $n = 100$ data points in dimension $d \in \{2, 4\}$, $\{x_1, \ldots, x_n\} \subseteq \mathbb{R}^{d+1}$ as follows. Each x_{ik} follows an independent and identically distributed Gaussian distribution with mean 0 and standard deviation 100. We fix $\boldsymbol{\beta}^t = (0, 1, \ldots, 1) \in \mathbb{R}^{d+1}$. The last coordinate, x_d, is chosen as the response and we generate it as:

$$x_{id} = -\sum_{k=1}^{d-1} x_{ik} + u_i, \qquad \forall i = 1, \ldots, n,$$

where u_i is also generated as a Gaussian distribution with mean 0 and standard deviation 10.

Then, 15% of the data are now corrupted by adding an extra Gaussian term (with mean 0 and standard deviation 500) to: (1) all the components except the last one or (2) to the last coordinate.

We get the fitting model for each one of the considered combinations (overall 42 models). Due to limitation of space in this paper, the complete results are available as a supplementary electronic material (see Appendix A). For each model we report: 1) the goodness of fitting index GoF, 2) the percentage of the sample data which are contained in a strip delimited by two parallel hyperplanes to $y = \widehat{\boldsymbol{\beta}}x$ with (orthogonal) distance $\varepsilon = 10$ (%), and 3) the width of the strip that is necessary to include 90% of the data (ϵ_{90}).

We conclude from these results that, in general, a better performance is observed in all the methods when the corrupted coordinate is the dependent one (Y), as compared with introducing the perturbation on the independent coordinate (X). In particular, the use of the SUM, the 1.5SUM and the kC criteria (for vertical distance residuals) empirically implies better models in the Y-corrupted case. Although slightly better, almost similar results were obtained for models based on AkC, MEDIAN and kC (for ℓ_τ residuals) due to their stability against extremal observations. Finally, we also point out that for the X-corrupted case, all models (except the AkC) coincide under the use of residuals measured by V, ℓ_1 and ℓ_∞. This is not the case for the results with Y-corrupted data, where equal or similar models were obtained for all the ℓ_τ-residuals.

Similar conclusions can be derived from the multivariate case ($d = 4$), except that in this situation there are no coincidences between the models obtained with different combinations of criteria and residuals. Furthermore, the convenience of using goodness of fitting measures which are not criterion/residual dependent is confirmed.

6.2. Data: Durbin–Watson

We also performed some experiments over the classical real data sample used in Durbin and Watson (1951). The data aims to analyze the annual consumption of spirits from 1870 to 1938 ($n = 69$) from the incomes and the relative price of spirits (deflated by a cost-of-living index). Hence, the variables observed in this data sets are the logarithms (the coefficients are then interpreted in terms of percent change) of the following measures: X_1 (Real income per head), X_2 (Relative price of spirits) and X_3 (Consumption of spirits per head).

For illustrative purposes, we analyze both the global model with the three variables ($d = 3$) and the bivariate model considering X_1 and X_3 and obviating X_2 ($d = 2$).

6.2.1. Bivariate model

For the case $d = 2$, the obtained hyperplanes are detailed in Table 4 and they are drawn in Fig. 6. Note that the methods that use vertical distance residuals (V) were not able to capture the actual behavior of the consumption with respect to the incomes. Fur-

Table 4
Estimations for the bidimensional Durbin–Watson's dataset.

	V	ℓ_1	ℓ_∞
SUM	(4.0898, −1.1454, −1)	(10.8840, −4.6184, −1)	(8.9764, −3.6797, −1)
MAX	(1.6986, −0.0196, −1)	(1.6986, −0.0196, −1)	(−0.5963, 1.1530, −1)
SOS	(2.9993, −0.6309, −1)	(13.5934, −6.0703, −1)	(7.0978, −2.7353, −1)
1.5SUM	(4.0730, −1.1566, −1)	(10.6113, −4.5067, −1)	(7.9926, −3.1851, −1)
kC	(5.5288, −1.9236, −1)	(8.7033, −3.5303, −1)	(7.6654, −2.9977, −1)
AkC	(2.7467, −0.4031, −1)	(17.1272, −7.6311, −1)	(18.4349, −8.2833, −1)
MED	(2.4167, −0.2310, −1)	(28.0156, −13.0469, −1)	(23.4462, −10.7748, −1)
	$\ell_{1.5}$	ℓ_2	ℓ_3
SUM	(10.8840, −4.6184, −1)	(10.8746, −4.6138, −1)	(9.8917, −4.1344, −1)
MAX	(1.6986, −0.0196, −1)	(−0.5963, 1.1530, −1)	(−0.5963, 1.1530, −1)
SOS	(13.1400, −5.8376, −1)	(10.9561, −4.7162, −1)	(8.7832, −3.6006, −1)
1.5SUM	(10.4466, −4.4233, −1)	(9.6868, −4.0399, −1)	(8.9821, −3.6851, −1)
kC	(8.0130, −3.1750, −1)	(8.0455, −3.1914, −1)	(8.5389, −3.4427, −1)
AkC	(13.9827, −6.0670, −1)	(21.0745, −9.6064, −1)	(20.6955, −9.4349, −1)
MED	(24.0656, −11.0819, −1)	(6.4510, −2.4601, −1)	(28.0150, −13.0466, −1)

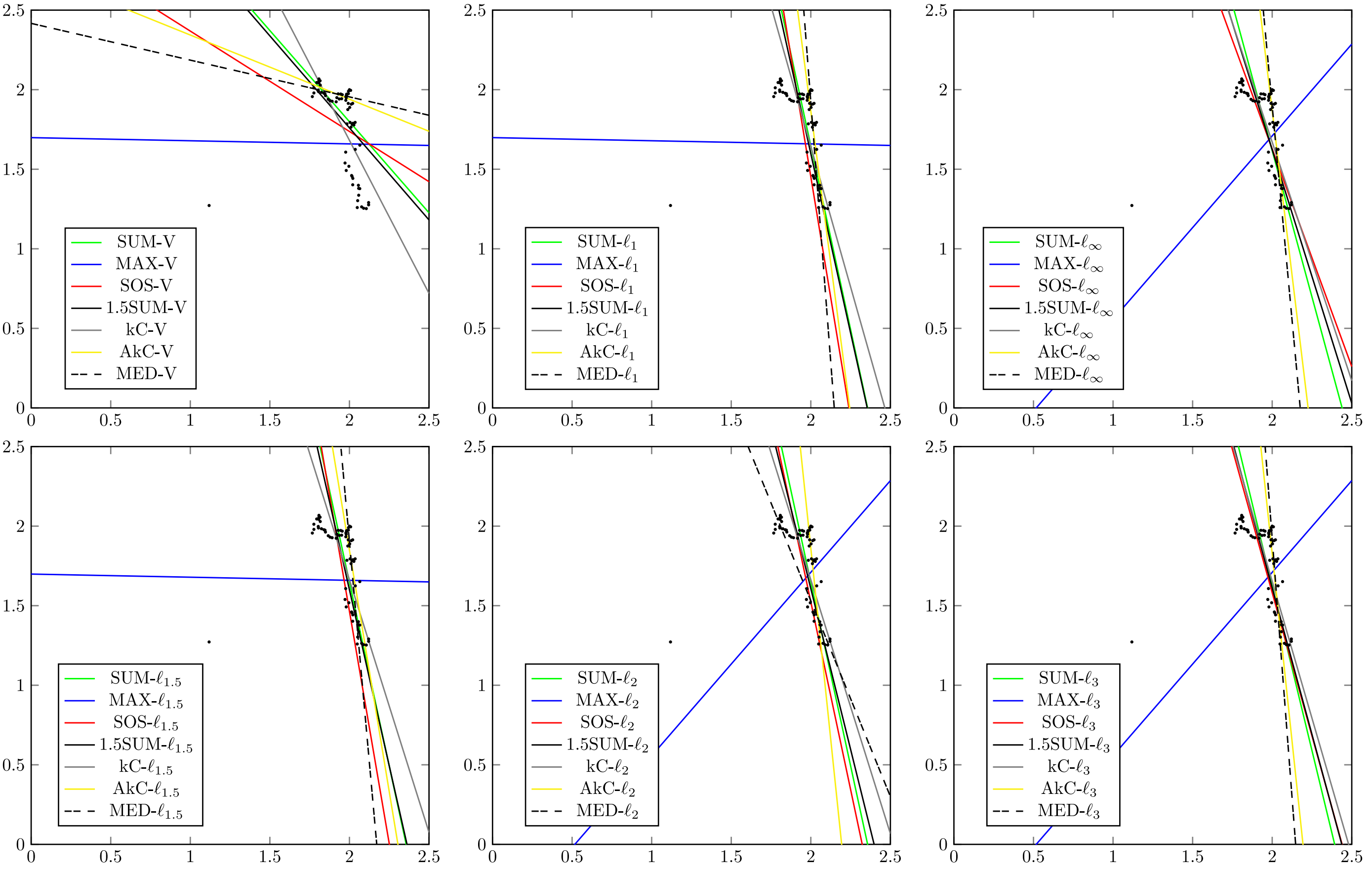


Fig. 6. Estimated lines for the data in Durbin and Watson (1951).

thermore, the MAX criterion seems to fail for any choice of residuals, since it tries to accommodate the unique outlier point that exists in the data set. The rest of the hyperplanes have a similar behavior. In order to analyze the differences between these models we also report, in Table 5, the marginal variations of each one of the models (according to Lemma 2.1).

Observe that, when the ℓ_1 residuals are considered, all except the MAX criterion provide a 0 marginal variation. This pattern can be explained as a result of Lemma 2.2 and the fact that the ℓ_1-norm unit ball in $\mathbb{R}^2$ has extreme points $\{\pm(0,\ 1),\ \pm(1,\ 0)\}$.

Table 5
Marginal variations for each of the models.

	V	ℓ_1	ℓ_∞	$\ell_{1.5}$	ℓ_2	ℓ_3
SUM	−1.1455	0	−0.7863	−0.0464	−0.2070	−0.4395
MAX	−0.0196	−0.0196	0.5355	−0.0196	0.4949	0.5151
SOS	−0.6309	0	−0.7322	−0.0291	−0.2029	−0.4597
1.5SUM	−1.1566	0	−0.7610	−0.0505	−0.2332	−0.4564
kC	−1.9236	0	−0.7498	−0.0961	−0.2853	−0.4660
AkC	−0.4032	0	−0.8922	−0.0270	−0.1029	−0.3147
MED	−0.2310	0	−0.9150	−0.0081	−0.3488	−0.2711

Table 6
Summary of k-fold cross validations experiments for the bidimensional Durbin–Watson's dataset.

		V	ℓ_1	ℓ_∞	$\ell_{1.5}$	ℓ_2	ℓ_3
SUM	min ε_{90}	0.1590	0.0560	0.0702	0.0491	0.0459	0.0560
	max ε_{90}	0.3049	0.1645	0.1444	0.1477	0.1480	0.1480
	medianε_{90}	0.2366	0.0983	0.0923	0.0881	0.0828	0.0983
	$\bar{\varepsilon}_{90}$	0.2330	0.1027	0.0982	0.0958	0.0959	0.1021
MAX	min ε_{90}	0.1262	0.1274	0.1262	0.1262	0.1262	0.1274
	max ε_{90}	0.3955	0.3955	0.3663	0.3663	0.3663	0.3955
	medianε_{90}	0.3664	0.3664	0.3621	0.3621	0.3621	0.3664
	$\bar{\varepsilon}_{90}$	0.3337	0.3338	0.3222	0.3222	0.3222	0.3338
SOS	min ε_{90}	0.1372	0.0844	0.0566	0.0568	0.0633	0.0793
	max ε_{90}	0.4072	0.1264	0.1163	0.1202	0.1235	0.1253
	medianε_{90}	0.2878	0.0962	0.0983	0.0879	0.0961	0.0961
	$\bar{\varepsilon}_{90}$	0.2980	0.1005	0.0973	0.0900	0.0905	0.0983
1.5SUM	min ε_{90}	0.1437	0.0476	0.0488	0.0524	0.0499	0.0478
	max ε_{90}	0.3091	0.1353	0.1199	0.1254	0.1308	0.1334
	medianε_{90}	0.2260	0.0834	0.0852	0.0910	0.0885	0.0841
	$\bar{\varepsilon}_{90}$	0.2349	0.0922	0.0872	0.0869	0.0884	0.0917
kC	min ε_{90}	0.1236	0.0414	0.0655	0.0495	0.0480	0.0412
	max ε_{90}	0.2843	0.1220	0.1147	0.1163	0.1185	0.1219
	medianε_{90}	0.1281	0.0837	0.0837	0.0851	0.0851	0.0855
	$\bar{\varepsilon}_{90}$	0.1511	0.0827	0.0834	0.0800	0.0809	0.0821
akC	min ε_{90}	0.4482	0.0421	0.0429	0.0367	0.0892	0.0484
	max ε_{90}	0.6677	0.2039	0.1853	0.2122	0.4654	0.1981
	medianε_{90}	0.5162	0.1722	0.1296	0.1605	0.1534	0.1466
	$\bar{\varepsilon}_{90}$	0.5282	0.1434	0.1338	0.1417	0.1914	0.1373
MED	min ε_{90}	0.4275	0.1182	0.1147	0.0979	0.1182	0.0615
	max ε_{90}	0.6375	0.2170	0.4612	0.2203	0.2137	0.2101
	medianε_{90}	0.5503	0.1712	0.1761	0.1701	0.1393	0.1565
	$\bar{\varepsilon}_{90}$	0.5406	0.1651	0.2093	0.1614	0.1501	0.1478

Hence,

$$\mathrm{k}(\beta) = \begin{cases} 1 & \text{if } \beta_3 = \max\{|\beta_1|, |\beta_3|\}, \\ -1 & \text{if } \beta_3 = -\max\{|\beta_1|, |\beta_3|\}, \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the marginal variation of X_1 with respect to X_3 is zero iff $|\beta_1| = \max\{|\beta_1|, |\beta_3|\}$, being then $|\beta_3| < |\beta_1|$. Observe that the latest implies that if the fitting line is rewritten in the form $X_3 = \gamma_0 + \gamma_1 X_1$, the absolute value of the slope of the line, $|\gamma_1|$, is greater than 1, being then the percent decreasing (or increasing) of the consumption (X_3) in term of the incomes (X_1), more than 100%.

In order to validate and analyze the stability of the computed hyperplanes we perform a *k*-fold cross validation scheme (Stone, 1974) to the data set. Such a method consists of randomly partitioning the sample into *k* folds of similar size, $S_1, \ldots, S_k$. For each $j \in \{1, \ldots, k\}$, each optimal hyperplane is computed using the points in $\bigcup_{i \neq j} S_i$ and S_j is used to validate the results. In our case, we partitioned the data into $k = 7$ folds, each of them with 10 data, except one with 9 points. In Table 6 we summarize the results obtained with this experiment. We report: the maximum, minimum, median and mean width of the strips that are necessary to cover the 90% of the (validation) data for the seven runs.

From the above results, we observe that the models that use vertical distance residuals need, in general, larger strips to cover the 90% of the points. The strips are particularly large for the MEDIAN criterion, where the widest strips were obtained. This conclusion is justified since the quantile criteria accommodate a single point, but do not take into account the deviations to the remaining elements in the data (apart from the ordering in the residuals). Also, for the same reason, the conservative MAX criterion makes the models to require wider strips. The residuals that produce the smallest range between the maximum and minimum length of the strips, are the ℓ_1, $\ell_{1.5}$, and ℓ_3; and for these type of residuals the *K*-centrum (*k*C) criterion gets the best results.

To illustrate the quality of the optimal hyperplanes, in Fig. 7 we show the values of the consumptions versus the actual consumptions for the first random fold in the experiments (in the validation sample that was not used to compute the hyperplanes).

The conclusions are that the models that use V and ℓ_∞-based residuals do not fit well to the actual trend of the validation data. The same conclusion also applies to the models that use the MAX criterion. On the other hand, all the models based on ℓ_τ-residual seem to fit quite-well to the data. As expected the *k*C and A*k*C criteria, which are known to be robust against extremal observations, actually capture the main information about the trend.

6.2.2. Complete models

We also performed the same experiments using all the variables: X_1 (incomes), X_2 (prices) and X_3 (consumptions). The optimal hyperplanes are shown in Table 7 (since the coefficients are non zero they were divided by $-\beta_3$ resulting in simplified models in the form $X_3 = \beta_0 + \beta_1 X_1 + \beta_2 X_2$.)

The summary of the results of the *k*-fold cross validation scheme (where the dataset was partitioned exactly as in the bivariate case) is shown in Table 8. Fig. 8 shows the values of the consumptions versus the actual consumptions for the first random fold in the experiments. From the results, one can observe that including all the variables in the model reduces the differences among the different methods. In this case, the consumption seems to be well linearly described by the incomes and prices. This conclusion is supported both by the projection and by the summary of *k*-cross validation experiments. The exceptionally bad performance of the MAX criterion in the bivariate case, is now as good as the rest of the criteria. In addition, the inclusion of prices in the model fixes the, in most cases, senseless signs of the coefficients in the bivariate models in Table 5. One can observe that in those cases an increase of the incomes would predict a decrease of the consumptions.

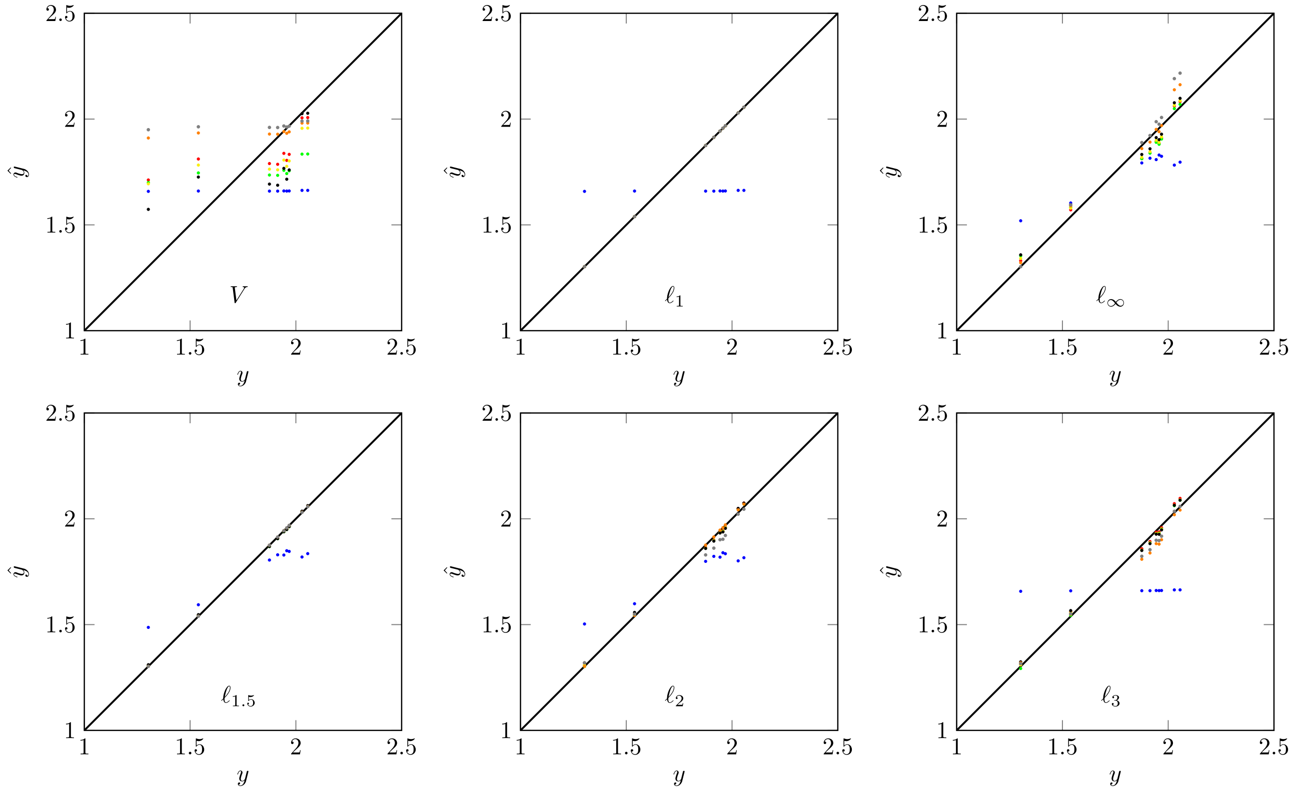


Fig. 7. Responses in the dependent variable by residuals for the bivariate case (SUM: red, MAX: blue, SOS: green, 1.5SUM: yellow, kC: black, AkC: orange, MEDIAN: gray). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 7
Estimations for the Durbin–Watson's dataset.

	V	ℓ_1	ℓ_∞
SUM	$(4.4817, 0.0696, -1.3374, -1)$	$(4.555, 0.0587, -1.3623, -1)$	$(4.1367, 0.3502, -1.4305, -1)$
MAX	$(4.5227, 0.0646, -1.3519, -1)$	$(4.6159, -0.013, -1.3273, -1)$	$(4.1355, 0.5086, -1.5758, -1)$
SOS	$(3.9725, 0.0331, -1.0692, -1)$	$(4.404, 0.1369, -1.3881, -1)$	$(4.404, 0.1369, -1.3881, -1)$
1.5SUM	$(4.404, 0.1369, -1.3881, -1)$	$(4.404, 0.1369, -1.3881, -1)$	$(4.404, 0.1369, -1.3881, -1)$
kC	$(4.4159, 0.0288, -1.2753, -1)$	$(4.4905, 0.0635, -1.3425, -1)$	$(4.3334, 0.1325, -1.3317, -1)$
AkC	$(4.4355, 0.0655, -1.3183, -1)$	$(4.4521, 0.0585, -1.3197, -1)$	$(4.4688, 0.0535, -1.323, -1)$
MED	$(4.4288, 0.0488, -1.2979, -1)$	$(4.5075, 0.0634, -1.3476, -1)$	$(4.3559, 0.1431, -1.3489, -1)$
	$\ell_{1.5}$	ℓ_2	ℓ_3
SUM	$(4.4445, 0.0698, -1.3242, -1)$	$(4.472, 0.0633, -1.331, -1)$	$(4.4922, 0.0619, -1.3386, -1)$
MAX	$(4.4155, 0.0352, -1.2797, -1)$	$(4.3938, 0.1107, -1.3377, -1)$	$(4.2655, 0.1691, -1.3326, -1)$
SOS	$(4.3498, 0.1131, -1.3201, -1)$	$(4.3498, 0.1131, -1.3201, -1)$	$(4.3498, 0.1131, -1.3201, -1)$
1.5SUM	$(4.2123, 0.4308, -1.5386, -1)$	$(4.0853, 0.4429, -1.4891, -1)$	$(3.6048, 0.7761, -1.5744, -1)$
kC	$(5.2647, -0.6758, -1.0312, -1)$	$(3.5719, 1.1094, -1.8642, -1)$	$(3.4912, 1.0623, -1.7796, -1)$
AkC	$(4.1061, 0.5015, -1.551, -1)$	$(4.1579, 0.467, -1.5434, -1)$	$(4.2963, 0.3239, -1.4761, -1)$
MED	$(4.3576, 0.2689, -1.4559, -1)$	$(4.0772, 0.4066, -1.4415, -1)$	$(76.3635, 25.0913, -61.4268, -1)$

6.3. Scalability

Finally, we would like to add some comments on the scalability of the proposed methods. As observed from the computational experiments, our formulations work well in the range of several hundreds of points regardless of the dimension of the space (within a reasonable limit). This is partly induced by the use of sortings in the aggregation criteria. Moving up to the range of thousands requires some further extensions by aggregation techniques (see Francis et al., 2000) that are beyond the scope of this manuscript. In spite of that, we have included an illustrative example with several thousands of points. Technical details on the accuracy of these techniques will be the subject of a forthcoming paper.

Example 6.1. We have randomly generated 2000 points in $\mathbb{R}^2$ with the same setting that in Subsection 6.1, by corrupting the last coordinate (X_2). The points are drawn in the right picture of Fig. 9 and are available at http://bit.ly/data2000. In order to show the scalability of the proposed methodology we have implemented a randomized aggregation technique based on Francis et al. (2000) to the computationally hardest models, i.e., those where the aggregation criterion is $\Phi \equiv$ AkC (with $k = \lfloor 0.5n \rfloor$) and residuals measured with vertical distance V, ℓ_1-norm and ℓ_2-norm. We report in Fig. 9 (left table) the estimated coefficients for the three models as well as the best objective values found and the computation times (in seconds) needed to obtain these solutions. As can be observed in Fig. 9 (right), the solutions that result with the aggregation technique have a good performance in terms of the geometric fitting. These techniques have been proved to find accurate solutions in reasonable computing times, so the models proposed in this paper are applicable to real-world datasets.

Table 8
Summary of k-fold cross validations experiments for the Durbin–Watson's dataset.

		V	ℓ_1	ℓ_∞	$\ell_{1.5}$	ℓ_2	ℓ_3
SUM	min ε_{90}	0.0369	0.0388	0.0315	0.0380	0.0346	0.0347
	max ε_{90}	0.0735	0.0741	0.0832	0.0743	0.0743	0.0732
	medianε_{90}	0.0629	0.0627	0.0647	0.0625	0.0625	0.0626
	ε_{90}	0.0573	0.0598	0.0616	0.0580	0.0567	0.0593
MAX	min ε_{90}	0.0562	0.0515	0.0515	0.0515	0.0515	0.0515
	max ε_{90}	0.0807	0.0762	0.0760	0.0760	0.0760	0.0762
	medianε_{90}	0.0701	0.0607	0.0644	0.0644	0.0607	0.0607
	ε_{90}	0.0678	0.0624	0.0641	0.0641	0.0624	0.0624
SOS	min ε_{90}	0.0255	0.0362	0.0310	0.0321	0.0327	0.0327
	max ε_{90}	0.0656	0.0683	0.0691	0.0678	0.0675	0.0675
	medianε_{90}	0.0586	0.0583	0.0568	0.0586	0.0581	0.0582
	ε_{90}	0.0547	0.0541	0.0537	0.0543	0.0528	0.0529
1.5SUM	min ε_{90}	0.0262	0.0342	0.0292	0.0308	0.0314	0.0316
	max ε_{90}	0.0685	0.0709	0.0713	0.0691	0.0703	0.0703
	medianε_{90}	0.0617	0.0563	0.0587	0.0559	0.0556	0.0558
	ε_{90}	0.0553	0.0547	0.0546	0.0527	0.0531	0.0532
kC	min ε_{90}	0.0269	0.0368	0.0265	0.0251	0.0272	0.0272
	max ε_{90}	0.0650	0.0700	0.0698	0.0709	0.0709	0.0700
	medianε_{90}	0.0588	0.0564	0.0559	0.0559	0.0569	0.0571
	ε_{90}	0.0514	0.0549	0.0536	0.0534	0.0538	0.0535
akC	min ε_{90}	0.0349	0.0338	0.0360	0.0305	0.0256	0.0604
	max ε_{90}	0.1042	0.1041	0.1017	0.3524	0.1100	0.1303
	medianε_{90}	0.0906	0.0888	0.0820	0.0885	0.0676	0.0931
	ε_{90}	0.0815	0.0799	0.0778	0.1115	0.0713	0.0923
MED	min ε_{90}	0.0342	0.0329	0.0346	0.0332	0.0429	0.0270
	max ε_{90}	0.1064	0.0994	0.0997	0.1102	0.3410	0.3266
	medianε_{90}	0.0709	0.0872	0.0894	0.0649	0.0844	0.0714
	ε_{90}	0.0738	0.0784	0.0794	0.0671	0.1215	0.1012

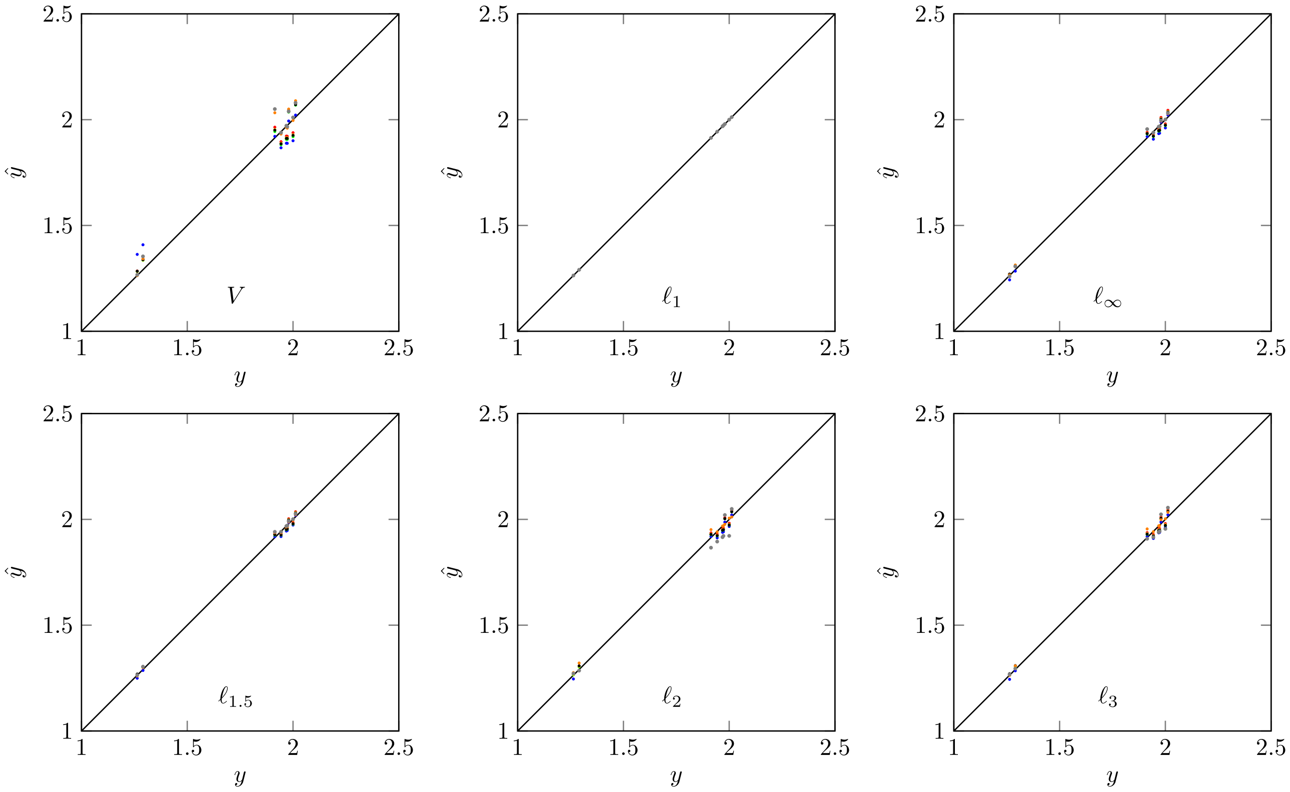


Fig. 8. Responses in the dependent variable by residuals for the $d = 3$ case (SUM: red, MAX: blue, SOS: green, 1.5SUM: yellow, kC: black, AkC: orange, MEDIAN: gray). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

7. Conclusions and further research

This paper generalizes previous attempts for modeling the problem of fitting hyperplanes to a given set of points. This approach allows for the combination of distance-based residuals aggregated by generalized ordered weighted averaging criteria. In addition, we provide unified mathematical programming formulations for all those models that allow one to use off-the-shelf solvers to handle the resulting problems. Two important particular cases of residuals are analyzed in more detail, namely those

	$\boldsymbol{\beta}$	Φ^*	CPUTime
V	$(0.4909, 0.9841, 1.0000)$	77541.55	214.816
ℓ_1	$(-0.6013, 1.0000, 1.0000)$	4343.721	175.463
ℓ_2	(-0.8166, 0.7103, 0.7038)	3041.330	1118.200

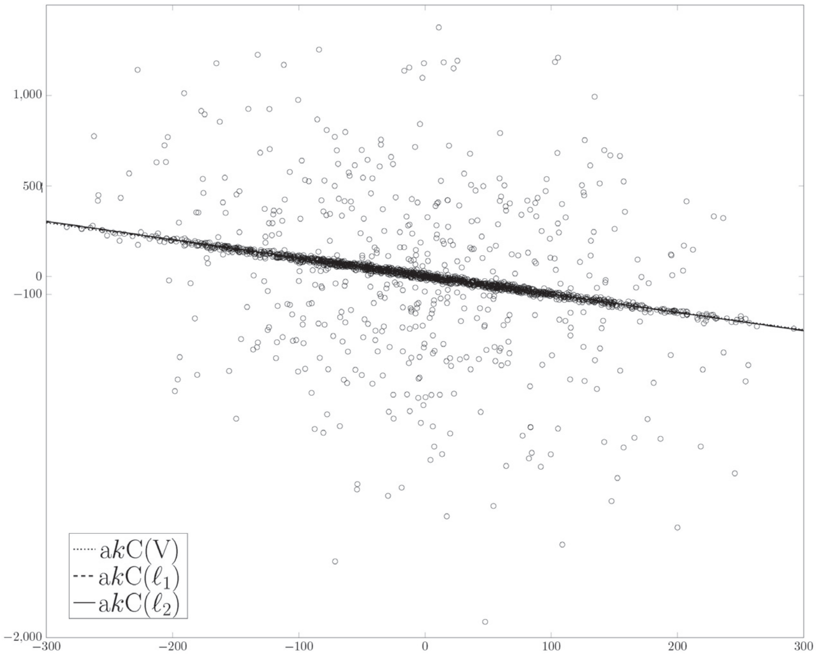


Fig. 9. Estimations for the instance of Example 6.1.

induced by block-and-ℓ_τ norms for $\tau \geq 1$. A new goodness of fitting measure is also introduced for this framework, which extends the classical coefficient of determination in least sum of squares fitting with vertical distances. Some illustrative computational experiments run in Gurobi under R are reported in order to illustrate and validate the new methodology for computing optimal fitting hyperplanes.

The results in this paper admit several extensions, still applying similar tools. Among them, we mention the study of the statistical analysis of the generalized noise terms, on the original data, that induce general norms residuals. In particular, we have conducted some preliminary tests to analyze the empirical distribution of hexagonal (see Example 4.5) and ℓ_2-norm based errors used in some of our computational experiments. We have compared whether the errors induced by the LSS criterion with the usual vertical distance and the sum criterion with the hex-and-ℓ_2-norms come from the same statistical distribution. Using the Mann–Whitney *U* test, to compare if two samples are identically distributed, we conclude that the three types of residuals come from the same distribution (the three null hypotheses cannot be rejected at a significance level of 5%). We have also raised the issue of regularization, i.e., adding constraints to overcome ill-posed data set, as well as the simultaneous computation of several (more than one) hyperplanes to a given data set such that each single point is “allocated” to its *closest model*, as in Bradley and Mangasarian (2000). Another interesting extension is the use of mathematical programming tools to fit hyperplanes to binary data. The usual techniques to estimate those models are based on likelihood estimation since least squares estimation is known to get poor results on this type of data. Here our proposal will fit in a natural way and will deserve further attention.

Acknowledgements

The authors would like to thank the handling editor and the anonymous referees for their constructive comments on previous versions of the paper. The first and second authors were partially supported by the project MTM2016-74983-C2-1-R and MTM2013-46962-C2-1-P (MINECO, Spain). The first and third authors were also supported by the research project PP2016-PIP06 (Universidad de Granada) and the research group SEJ-534 (Junta de Andalucía).

Appendix A. Supplementary electronic material

See Tables A.9–A.12.

Table A.9
Results for bidimensional experiments corrupting the X variables.

		V	ℓ_1	ℓ_∞
SUM	$\widehat{\boldsymbol{\beta}}$	$(-1.9587, 0.3011, 1)$	$(1.9587, -0.3011, -1)$	$(0.4240, -0.9403, -1)$
	GoF	0.1456	0.1456	0.5342
	%	8%	8%	65%
	ϵ_{90}	141.2995	141.2995	87.0871
MAX	$\widehat{\boldsymbol{\beta}}$	$(10.9038, 0.1571, 1)$	$(10.9038, 0.1571, 1)$	$(10.9038, 0.1571, 1)$
	GoF	0.1484	0.1484	0.2641
	%	10%	10%	10%
	ϵ_{90}	158.9295	158.9295	158.9295
SOS	$\widehat{\boldsymbol{\beta}}$	$(-3.1753, 0.1860, 1)$	$(3.1753, -0.1860, -1)$	$(-1.8549, 0.2858, 1)$
	GoF	0.2261	0.2261	0.4925
	%	8%	8%	9%
	ϵ_{90}	157.7177	157.7177	143.1279
1.5SUM	$\widehat{\boldsymbol{\beta}}$	$(-3.5386, 0.2112, 1)$	$(3.5397, -0.2112, -1)$	$(0.3967, -0.4136, -1)$
	GoF	0.1812	0.1812	0.4499
	%	8%	8%	8%
	ϵ_{90}	152.361	152.3626	127.4389
kC	$\widehat{\boldsymbol{\beta}}$	$(-3.0188, 0.2328, 1)$	$(-3.0188, 0.2328, 1)$	$(0.3503, 0.9091, 1)$
	GoF	0.1226	0.1226	0.4275
	%	8%	8%	60%
	ϵ_{90}	150.5599	150.5599	85.1974
AkC	$\widehat{\boldsymbol{\beta}}$	$(5.8180, 0.7718, 1)$	$(2.2956, 0.7734, 1)$	$(2.6795, 0.9874, 1)$
	GoF	0.6735	0.9040	0.9758
	%	29%	34%	70%
	ϵ_{90}	77.4723	74.8420	92.8187
MED	$\widehat{\boldsymbol{\beta}}$	$(6.1846, 0.7795, 1)$	$(6.1842, 0.7795, 1)$	$(1.3314, 0.9890, 1)$
	GoF	0.7021	0.8690	0.9741
	%	31%	31%	70%
	ϵ_{90}	78.4775	78.4772	91.9773
		$\ell_{1.5}$	ℓ_2	ℓ_3
SUM	$\widehat{\boldsymbol{\beta}}$	$(-0.2603, -0.9299, -1)$	$(-0.2603, -0.9299, -1)$	$(-0.2603, -0.9299, -1)$
	GoF	0.4133	0.3417	0.2615
	%	62%	62%	62%
	ϵ_{90}	86.7791	86.7791	86.7791
MAX	$\widehat{\boldsymbol{\beta}}$	$(-10.9038, -0.1571, -1)$	$(-10.9038, -0.1571, -1)$	$(10.9038, 0.1571, 1)$
	GoF	0.1821	0.1588	0.1495
	%	10%	10%	10%
	ϵ_{90}	158.9295	158.9295	158.9295
SOS	$\widehat{\boldsymbol{\beta}}$	$(2.4728, -0.2391, -1)$	$(-2.8551, 0.2102, 1)$	$(-3.1181, 0.1903, 1)$
	GoF	0.3163	0.2552	0.2295
	%	8%	8%	8%
	ϵ_{90}	149.8204	151.9362	156.6873
1.5SUM	$\widehat{\boldsymbol{\beta}}$	$(3.4138, -0.2225, -1)$	$(3.0670, -0.2704, -1)$	$(1.4864, -0.3260, -1)$
	GoF	0.1853	0.2145	0.2799
	%	8%	9%	7%
	ϵ_{90}	149.6913	145.969	135.7776
kC	$\widehat{\boldsymbol{\beta}}$	$(-2.6422, 0.2474, 1)$	$(-0.2632, -0.9011, -1)$	$(-0.3503, -0.9091, -1)$
	GoF	0.1263	0.1913	0.2791
	%	9%	57%	60%
	ϵ_{90}	147.9623	84.4867	85.1974
AkC	$\widehat{\boldsymbol{\beta}}$	$(-0.0741, 0.9357, 1)$	$(2.2028, 1.0126, 1)$	$(-0.9506, 0.9930, 1)$
	GoF	0.9468	0.9576	0.9645
	%	64%	70%	65%
	ϵ_{90}	86.9840	94.2569	91.5147
MED	$\widehat{\boldsymbol{\beta}}$	$(1.5779, -0.9545, -1)$	$(2.9207, 1.0139, 1)$	$(0.2899, 0.9792, 1)$
	GoF	0.9530	0.9611	0.9655
	%	63%	69%	65%
	ϵ_{90}	88.5178	94.8548	90.5271

Table A.10
Results for bidimensional experiments corrupting the *Y* variables.

		V	ℓ_1	ℓ_∞
SUM	$\widehat{\beta}$	$(-0.4324, -1.0070, -1)$	$(-2.7476, -1.1156, -1)$	$(-0.8817, -1.0333, -1)$
	GoF	0.5226	0.5464	0.7637
	%	72%	57%	73%
	ϵ_{90}	158.3495	144.4862	154.9621
MAX	$\widehat{\beta}$	$(164.40, 1.95, -1)$	$(-131.52, -7.30, -1)$	$(-131.52, -7.30, -1)$
	GoF	0.0109	0.7575	0.7867
	%	5%	6%	6%
	ϵ_{90}	266.337	144.6019	144.6019
SOS	$\widehat{\beta}$	$(-19.4780, 0.9765, 1)$	$(24.3778, -3.9704, -1)$	$(-21.8989, 2.4558, 1)$
	GoF	0.2459	0.8055	0.8896
	%	24%	12%	14%
	ϵ_{90}	176.2108	119.0515	108.3728
1.5SUM	$\widehat{\beta}$	$(2.2257, -0.9993, -1)$	$(8.1241, -2.8635, -1)$	$(4.2013, -1.5531, -1)$
	GoF	0.3894	0.6583	0.8111
	%	72%	15%	24%
	ϵ_{90}	161.1331	114.1084	107.9904
kC	$\widehat{\beta}$	$(-0.6995, -0.9989, -1)$	$(4.8095, -1.6540, -1)$	$(-1.0107, -1.0744, -1)$
	GoF	0.4422	0.4969	0.7265
	%	71%	23%	67%
	ϵ_{90}	159.1129	100.6695	150.2014
AkC	$\widehat{\beta}$	$(10.0084, -0.9838, -1)$	$(-1.3062, -1.0398, -1)$	$(-1.2815, -0.9942, -1)$
	GoF	0.7526	0.9914	0.9961
	%	53%	70%	72%
	ϵ_{90}	168.5344	153.9189	159.2534
MED	$\widehat{\beta}$	$(8.6545, -0.9641, -1)$	$(-0.8028, -1.0379, -1)$	$(-4.3252, -1.0113, -1)$
	GoF	0.8478	0.9894	0.9947
	%	57%	73%	69%
	ϵ_{90}	170.0131	154.4849	155.1026
		$\ell_{1.5}$	ℓ_2	ℓ_3
SUM	$\widehat{\beta}$	$(-0.9890, -1.0403, -1)$	$(-0.9890, -1.0403, -1)$	$(-0.9890, -1.0403, -1)$
	GoF	0.6250	0.6658	0.7023
	%	70%	70%	70%
	ϵ_{90}	154.0857	154.0857	154.0857
MAX	$\widehat{\beta}$	$(-131.52, -7.30, -1)$	$(-131.52, -7.30, -1)$	$(-131.52, -7.30, -1)$
	GoF	0.7577	0.7598	0.7654
	%	6%	6%	6%
	ϵ_{90}	144.6019	144.6019	144.6019
SOS	$\widehat{\beta}$	$(24.0474, -3.7686, -1)$	$(23.2040, -3.2532, -1)$	$(22.5246, -2.8381, -1)$
	GoF	0.8077	0.8195	0.8412
	%	13%	13%	13%
	ϵ_{90}	118.4519	119.827	115.0321
1.5SUM	$\widehat{\beta}$	$(8.2797, -2.4830, -1)$	$(5.8395, -1.9194, -1)$	$(4.7010, -1.6953, -1)$
	GoF	0.6667	0.6976	0.7384
	%	14%	19%	23%
	ϵ_{90}	114.0191	102.4955	97.65193
kC	$\widehat{\beta}$	$(-1.0107, -1.0744, -1)$	$(-1.0107, -1.0744, -1)$	$(-0.8903, -1.0744, -1)$
	GoF	0.5665	0.6135	0.6556
	%	67%	67%	66%
	ϵ_{90}	150.2014	150.2014	150.2834
AkC	$\widehat{\beta}$	$(-2.6754, -1.0658, -1)$	$(-2.7011, -0.9640, -1)$	$(-3.9149, -1.0070, -1)$
	GoF	0.9901	0.9910	0.9915
	%	69%	68%	69%
	ϵ_{90}	150.0206	161.8515	155.8964
MED	$\widehat{\beta}$	$(-0.8019, -1.0319, -1)$	$(-2.6799, -1.0009, -1)$	$(-1.5141, -1.0345, -1)$
	GoF	0.9911	0.9924	0.9928
	%	74%	70%	70%
	ϵ_{90}	155.184	157.4707	154.3846

Table A.11
Results for experiments for $d = 4$ and corrupting the X variables.

		V	ℓ_1	ℓ_∞
SUM	$\widehat{\boldsymbol{\beta}}$	(8.7754, 0.2361, 0.1242, −0.0645, 1)	(−167.9861, 32.8678, −11.1472, −15.3593, 1)	(19.6624, 1.9411, 1.4336, −2.6949, 1)
	GoF	0.0369	0.3527	0.7030
	%	8%	9%	15%
	ϵ_{90}	285.1339	172.616	166.2396
MAX	$\widehat{\boldsymbol{\beta}}$	(11.2676, −0.8055, 0.4093, 0.3802, 1)	(95.4943, −2.3074, −2.7088, 4.5984, 1)	(76.9688, −2.1455, −2.9597, 4.6480, 1)
	GoF	0.1200	0.5037	0.7852
	%	2%	9%	6%
	ϵ_{90}	243.9038	160.86	164.3572
SOS	$\widehat{\boldsymbol{\beta}}$	(2.7637, 0.1306, 0.06391, −0.0111, 1)	(−35.0079, −17.4180, 5.1138, 8.8243, −1)	(14.4492, 2.3985, 1.8254, −3.4712, 1)
	GoF	0.0409	0.5787	0.9085
	%	6%	9%	8%
	ϵ_{90}	285.0815	170.37	165.6255
1.5SUM	$\widehat{\boldsymbol{\beta}}$	(3.1382, 0.1714, 0.0663, −0.03521)	(21.9152, −18.9245, 5.5144, 9.6284, −1)	(−20.1562, −2.0728, −1.5407, 2.9444, −1)
	GoF	0.0418	0.4776	0.8349
	%	7%	8%	14%
	ϵ_{90}	282.7383	167.7096	165.9725
kC	$\widehat{\boldsymbol{\beta}}$	(−6.8937, 0.1108, 0.0744, −0.0183, 1)	(−34.1432, −15.4977, 4.3066, 7.9523, −1)	(5.0421, 2.0898, 1.4381, −2.8638, 1)
	GoF	0.0258	0.3487	0.6984
	%	8%	8%	15%
	ϵ_{90}	276.4327	168.3023	169.65
AkC	$\widehat{\boldsymbol{\beta}}$	(−29.5486, 0.5489, 0.2119, 0.2342, 1)	(11.5813, 2.8055, −0.1579, 0.1805, 1)	(2.7269, 1.0225, 0.9985, 1.0072, 1)
	GoF	0.1544	0.8716	0.9950
	%	12%	5%	82%
	ϵ_{90}	304.1316	306.9669	496.6216
MED	$\widehat{\boldsymbol{\beta}}$	(11.3163, 0.5095, 0.5018, 0.0667, 1)	(15.2913, −1.38181, −0.1062, 9.6624, 1)	(2.3001, 1.0447, 1.0149, 1.0033, 1)
	GoF	0.3706	0.8308	0.9941
	%	9%	11%	80%
	ϵ_{90}	283.331	251.5948	497.3323
		$\ell_{1.5}$	ℓ_2	ℓ_3
SUM	$\widehat{\boldsymbol{\beta}}$	(−25.3339, 7.2803, 0.3850, −6.5208, 1)	(−25.3339, 7.2803, 0.3850, −6.5208, 1)	(−48.9741, −2.5251, −1.5173, 3.4889, −1)
	GoF	0.3973	0.4630	0.5446
	%	12%	12%	11%
	ϵ_{90}	167.1534	167.1534	163.8287
MAX	$\widehat{\boldsymbol{\beta}}$	(−76.9688, 2.1455, 2.9597, −4.6480, −1)	(−76.9688, 2.1455, 2.9597, −4.6480, −1)	(−76.9688, 2.1455, 2.9597, −4.6480, −1)
	GoF	0.5510345	0.6096547	0.677138
	%	6%	6%	6%
	ϵ_{90}	164.3572	164.3572	164.3572
SOS	$\widehat{\boldsymbol{\beta}}$	(−19.8365, −24.1780, −1.6843, 23.0309, −1)	(−37.1798, −20.6518, −4.8914, 22.4924, −1)	(16.2930, 4.1351, 2.2042, −5.3890, 1)
	GoF	0.6391	0.7149	0.7921
	%	9%	9%	4%
	ϵ_{90}	159.013	160.1321	165.3201
1.5SUM	$\widehat{\boldsymbol{\beta}}$	(27.4692, 14.0582, 1.0081, −12.9659, 1)	(27.4555, 14.0608, 1.0082, −12.9683, 1)	(−20.4048, −3.2308, −1.6763, 4.1796, −1)
	GoF	0.5314	0.6059	0.6909
	%	10%	10%	5%
	ϵ_{90}	162.8882	162.8875	164.1443
kC	$\widehat{\boldsymbol{\beta}}$	(31.8219, 41.5015, −5.2288, −30.4070, 1)	(2.4227, 14.3655, 4.4768, −15.4827, 1)	(6.6713, −3.7849, −1.5627, 4.3751, −1)
	GoF	0.3916	0.4629	0.5440
	%	5%	7%	4%
	ϵ_{90}	165.793	168.1855	165.9668
AkC	$\widehat{\boldsymbol{\beta}}$	(7.9530, −1.6065, 0.3482, 0.8960, −1)	(−25.2618, −1.0371, −1.4553, 0.7368, −1)	(40.7617, −1.6662, −0.5106, 0.5624, −1)
	GoF	0.7403	0.8148	0.8817
	%	7%	11%	9%
	ϵ_{90}	180.9401	244.0442	231.9954
MED	$\widehat{\boldsymbol{\beta}}$	(−28.1536, −1.9062, −0.5785, 0.5246, −1)	(−51.5261, 1.9897, 1.0285, −0.5282, 1)	(6.9522, 1.2873, 1.0511, −0.1044, 1)
	GoF	0.8278	0.8575	0.8941
	%	9%	8%	14%
	ϵ_{90}	237.8898	305.539	350.0691

Table A.12
Results for experiments for $d = 4$ and corrupting the Y variables.

		V	ℓ_1	ℓ_∞
SUM	$\widehat{\boldsymbol{\beta}}$	(1.9468, 0.9648, 0.9899, 1.0058, 1)	(−1.9158, −1.1083, −0.8751, −3.3186, −1)	(1.6655, −1.0083, −1.0530, −1.0446, −1)
	GoF	0.5999	0.6538	0.9006
	%	78%	14%	76%
	ϵ_{90}	123.5456	149.6274	121.8106
MAX	$\widehat{\boldsymbol{\beta}}$	(1 − 04.7766, −1.0780, −2.8506, −0.8355, −1)	(120.6153, −1.4207, −5.5268, −0.7782, −1)	(54.3395, 2.3207, 6.0411, 3.4977, 1)
	GoF	0.3357	0.8267	0.9078
	%	12%	7%	12%
	ϵ_{90}	151.6067	147.4952	138.4277
SOS	$\widehat{\boldsymbol{\beta}}$	(−12.1432, −0.8507, −1.0758, −1.1049, −1)	(25.1165, −1.2149, −5.4326, −1.1199, −1)	(−5.4787, −1.8048, −2.3397, −2.0389, −1)
	GoF	0.4247	0.9015	0.9801
	%	45%	13%	15%
	ϵ_{90}	124.0456	135.9287	102.1587
1.5SUM	$\widehat{\boldsymbol{\beta}}$	(−2.1265, −0.9557, −0.9984, −1.0235, −1)	(34.3751, −1.0783, −5.2458, −1.0619, −1)	(−0.6651, −1.3869, −1.5549, −1.5790, −1)
	GoF	0.5106	0.8044	0.9485
	%	77%	11%	22%
	ϵ_{90}	124.3694	139.4734	95.54551
kC	$\widehat{\boldsymbol{\beta}}$	(−0.3095, −0.9816, −1.0017, −1.009643, −1)	(2.1980, −0.8680, −0.9950, −3.4086, −1)	(−0.6929, −1.0211, −1.0606, −1.0666, −1)
	GoF	0.5275	0.6525	0.8835
	%	80%	10%	74%
	ϵ_{90}	123.0891	145.6142	120.8033
AkC	$\widehat{\boldsymbol{\beta}}$	(−7.2126, −0.9981, −1.2345, −0.9988, −1)	(−1.7307, −0.9801, −1.0396, −1.0121, −1)	(0.1128, −0.9847, −1.0149, −1.0013, −1)
	GoF	0.8785	0.9933	0.9981
	%	57%	77%	80%
	ϵ_{90}	105.7586	120.4785	121.9634
MED	$\widehat{\boldsymbol{\beta}}$	(−8.4437, −1.0328, −1.1891, −0.9958, −1)	(−3.0605, −0.9660 − 1.0175, −1.0366, −1)	(−1.7471, −0.9713, −0.9881, −1.0144, −1)
	GoF	0.9011	0.9921	0.9980
	%	58%	76%	79%
	ϵ_{90}	105.9371	123.0289	123.8959
		$\ell_{1.5}$	ℓ_2	ℓ_3
SUM	$\widehat{\boldsymbol{\beta}}$	(0.5934, −1.0202, −1.0588, −1.0264, −1)	(0.6616, −1.0203, −1.0584, −1.0270, −1)	(0.9775, −1.0098, −1.0563, −1.0343, −1)
	GoF	0.7489	0.8006	0.8418
	%	80%	80%	78%
	ϵ_{90}	119.4431	119.5293	120.6788
MAX	$\widehat{\boldsymbol{\beta}}$	(120.6153, −1.4207, −5.5268, −0.7782, −1)	(−54.3395, −2.3207, −6.0411, −3.4977, −1)	(−54.3395, −2.3207, −6.0411, −3.4977, −1)
	GoF	0.8267	0.8384	0.8643
	%	7%	12%	12%
	ϵ_{90}	147.4952	138.4277	138.4277
SOS	$\widehat{\boldsymbol{\beta}}$	(−14.4853, 1.5436, 4.4201, 1.5950, 1)	(−0.3904, 1.7361, 2.9264, 2.0617, 1)	(4.7620, 1.9721, 2.5444, 2.0415, 1)
	GoF	0.9022	0.9272	0.9514
	%	13%	10%	12%
	ϵ_{90}	131.3351	114.7621	106.4697
1.5SUM	$\widehat{\boldsymbol{\beta}}$	(15.7120, −1.1641, −2.6186, −1.8366, −1)	(−0.8627, −1.4497, −1.6239, −1.9098, −1)	(−0.6434, −1.4056, −1.5798, −1.5348, −1)
	GoF	0.8079	0.8565	0.8965
	%	21%	22%	20%
	ϵ_{90}	114.939	97.67539	97.29497
kC	$\widehat{\boldsymbol{\beta}}$	(−1.0976, −1.0234, −1.0643, −1.0656, −1)	(−1.0942, −1.0234, −1.0641, −1.0656, −1)	(−0.7613, −1.0216, −1.0617, −1.0665, −1)
	GoF	0.7053	0.7661	0.8144
	%	74%	74%	74%
	ϵ_{90}	120.25	120.262	120.6901
AkC	$\widehat{\boldsymbol{\beta}}$	(0.8072, −0.9319, −1.1111, −1.0901, −1)	(−1.5573, −0.9672, −0.9991, −1.0184, −1)	(2.4443, −1.0165, −0.9923, −1.0147, −1)
	GoF	0.9929	0.9954	0.9930
	%	64%	77%	82%
	ϵ_{90}	124.0139	123.7847	123.5452
MED	$\widehat{\boldsymbol{\beta}}$	(−0.6735, −0.9887, −1.0180, −0.9497, −1)	(0.4156, −0.9995, −1.0147, −1.0116, −1)	(−1.1572, −0.9753, −1.0309, −0.9853, −1)
	GoF	0.9945	0.9949	0.9964
	%	75%	81%	78%
	ϵ_{90}	118.3319	121.9701	120.0091

References

Amaldi, E., Coniglio, S., Taccari, L., 2016. Discrete optimization methods to fit piecewise affine models to data points. Comput. Oper. Res. 75, 214–230.

Atkinson, A.C., Cheng, T.C., 1999. Computing least trimmed squares regression with the forward search. Stat. Comput. 9, 251–263.

Balas, E., 1979. Disjunctive programming. Ann. Discrete Math. 5, 3–51.

Bargiela, A., Hartley, J.K., 1993. Orthogonal linear regression algorithm based on augmented matrix formulation. Comput. Oper. Res. 20 (8), 829–836.

Bertsimas, D., King, A., Mazumder, R., 2016. Best subset selection via a modern optimization lens. Ann. Stat. 44 (2), 813–852.

Bertsimas, D., Mazumder, R., 2014. Least quantile regression via modern optimization. Ann. Stat. 42 (6), 2494–2525.

Bertsimas, D., Shioda, R., 2007. Classification and regression via integer optimization. Oper. Res. 55 (2), 252–271.

Blanco, V., Puerto, J., El-Haj Ben-Ali, S., 2014. Revisiting several problems and algorithms in continuous location with ℓ_τ norms. Comput. Optim. Appl. 58 (3), 563–595.

Blanco, V., Puerto, J., Salmerón, R., 2016. A general framework for locating hyperplanes to fitting set of points. available at. https://arxiv.org/abs/1505.03451.

Boggs, P.T., Rogers, J.E., 1990. Orthogonal distance regression. Contemp. Math. 112, 183–194.

Bradley, P.S., Mangasarian, O.L., 2000. K-plane clustering. J. Global Opt. 16 (1), 23–32.

Carrizosa, E., Conde, E., Fernández, F.R., Muñoz, M., Puerto, J., 1995. Pareto optimality in linear regression. J. Math. Anal. Appl. 190, 129–141.

Carrizosa, E., Plastria, F., 1995. The determination of a “least quantile of squares regression line” for all quantiles. Comput. Stat. Data Anal. 20 (5), 467–479.

Cavalier, T., Melloy, B., 1991. An iterative linear programming solution to the Euclidean regression model. Comput. Oper. Res. 18 (8), 655–661.

Diaz-Báñez, J.M., Mesa, J.A., Schöbel, A., 2004. Continuous location of dimensional structures. Eur. J. of Oper. Res 152 (1), 22–44.

Drezner, Z., Steiner, S., Wesolowsky, G.O., 2002. On the circle closest to a set of points. Comput. Oper. Res 29 (6), 637–650.

Durbin, J., Watson, G.S., 1951. Testing for serial correlation in least squares regression II. Biometrika 38, 159–178.

Edgeworth, F.Y., 1887. On observations relating to several quantities. Hermathena 6, 279–285.

Fernández, E., Pozo, M.A., Puerto, J., 2014. Ordered weighted average combinatorial optimization: formulations and their properties.. Discrete Appl. Math. 169, 97–118.

Fernández, E., Pozo, M.A., Puerto, J., Scozzari, A., 2017. Ordered weighted average optimization in multiobjective spanning tree problems. Eur. J. Oper. Res. 260 (886903), 886–903.

Fletcher, P.T., 2013. Geodesic regression and the theory of least squares on Riemannian manifolds. Int. J. Comput. Vis. 105, 171–185.

Francis, R.L., Lowe, T.J., Tamir, A., 2000. Aggregation error bounds for a class of location models. Oper. Res. 48 (2), 294–307.

Gauss, C.F., 1809. Theoria Motus Corporum Coelestium in Sectionibus Conicis Solum Ambientium (Theory of the Motion of the Heavenly Bodies Moving about the Sun in Conic Sections). Dover Publications, Mineola.

Giloni, A., Padberg, M., 2002. Alternative methods of linear regression. Math. Comput. Model. 35 (3–4), 361–374.

Grzybowski, J., Nickel, S., Pallaschke, D., Urbański, R., 2011. Ordered median functions and symmetries. Optimization 60, 801–811.

Hampel, F.R., 1975. Beyond location parameters: robust concepts and methods. Bull. Int. Stat. Inst. 46, 375–382.

Hoerl, A., Kennard, R., 1988. Ridge regression. In: Encyclopedia of Statistical Sciences, vol. 8. Wiley, New York, pp. 129–136.

Hofmann, M., Gatu, C., Kontoghiorghes, E.J., 2010. An exact least trimmed squares algorithm for a range of coverage values.. J. Comput. Graph Stat. 19 (1), 191–204.

Humphreys, R.M., 1978. Studies of luminous stars in nearby galaxies. I. Supergiants and o stars in the milky way. Astrophys. J. Suppl. S. 38, 309–350.

Lee, S., Grossmann, I., 2000. New algorithms for nonlinear generalized disjunctive programming. Comput. Chem. Eng. 24, 2125–2214.

Love, R.F., Morris, J.G., 1972. Modelling inter-city road distances by mathematical functions. Oper. Res. Q 23 (1), 61–71.

Mangasarian, O.L., 1999. Arbitrary-norm separating plane. Oper. Res. Lett. 24 (1–2), 15–23.

Marín, A., Nickel, S., Puerto, J., Velten, S., 2009. A flexible model and efficient solution strategies for discrete location problems. Discrete Appl. Math. 157 (5), 1128–1145.

McKean, J.W., Sievers, G.L., 1987. Coefficients of determination for least absolute deviation analysis. Stat. Probab. Lett. 5 (1), 49–54.

Megiddo, N., Tamir, A., 1983. Finding least-distance lines. SIAM J. Algebraic Discrete Methods 4 (2), 207–211.

Miller, A., 2002. Subset Selection in Regression. CRC Press, Washington.

Miyashiro, R., Takano, Y., 2015. Mixed integer second-order cone programming formulations for variable selection in linear regression. Eur. J. Oper. Res. 247 (3), 721–731.

Narula, S.C., Wellington, J.F., 2007. Multiple criteria linear regression. Eur. J. Oper. Res. 181 (2), 767–772.

Nickel, S., Puerto, J., 1999. A unified approach to network location. Networks 34, 283–290.

Nickel, S., Puerto, J., 2005. Location Theory: A Unified Approach. Springer Verlag.

Pinson, P., Nielsen, H., Madsen, H., Nielsen, T., 2008. Local linear regression with adaptive orthogonal fitting for the wind power application. Stat. Comput 58 (1), 59–71.

Rousseeuw, P., 1984. Least median of squares regression. J. Am. Stat. Assoc. 79, 871–880.

Rousseeuw, P., Leroy, A., 2003. Robust Regression and Outlier Detection. Wiley, New York.

Rousseeuw, P.J., 1983. Multivariate estimation with high breakdown point. Math. Stat. App. B 283–297.

Schöbel, A., 1996. Locating least-distant lines with block norms. Stud. Locat. Anal. 10, 139–150.

Schöbel, A., 1997. Locating line segments with vertical distances. Stud. Locat. Anal. 11, 143–158.

Schöbel, A., 1998. Locating least distant lines in the plane. Eur. J. Oper. Res 106 (1), 152–159.

Schöbel, A., 1999. Locating Lines and Hyperplanes: Theory and Algorithms. Kluwer Academic Publishers.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. B 36, 111–147.

Thoai, R., 1999. D.C. programming: an overview. J. Optimiz. Theory Appl. 193 (1), 1–43.

Van Huffel, S., Vanderwalle, J., 1991. The total least squares problem: computational aspects and analysis. SIAM Front. Appl. Math. ISBN: 978-0-89871-275-9.

Ward, J.E., Wendell, R.E., 1980. A new norm for measuring distance which yields linear location models. Oper. Res 28, 836–844.

Ward, J.E., Wendell, R.E., 1985. Using block norms for location modeling. Oper. Res. 33, 1074–1090.

Yager, R.R., Beliakov, G., 2010. OWA operators in regression problems. IEEE Trans. Fuzzy Syst. 18 (1), 106–113.