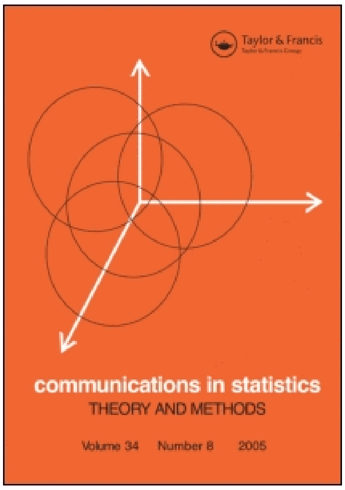

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597238

A Generalized Regression Methodology for Bivariate Heteroscedastic Data

Antonio Fernández[a]; Manuel Vázquez[ab]
[a] Electronica Fisica, EUITT-UPM, Madrid, Spain [b] IES, UPM, Madrid, Spain

Online publication date: 02 December 2010

To cite this Article Fernández, Antonio and Vázquez, Manuel(2011) 'A Generalized Regression Methodology for Bivariate Heteroscedastic Data', Communications in Statistics - Theory and Methods, 40: 4, 598 — 621
To link to this Article: DOI: 10.1080/03610920903444011
URL: http://dx.doi.org/10.1080/03610920903444011

Communications in Statistics—Theory and Methods, 40: 598–621, 2011
Copyright © Taylor & Francis Group, LLC
ISSN: 0361-0926 print/1532-415X online
DOI: 10.1080/03610920903444011

A Generalized Regression Methodology for Bivariate Heteroscedastic Data

ANTONIO FERNÁNDEZ[1] AND MANUEL VÁZQUEZ[1,2]

[1]Electronica Fisica, EUITT-UPM, Madrid, Spain
[2]IES, UPM, Madrid, Spain

We present a methodology for reducing a straight line fitting regression problem to a Least Squares minimization one. This is accomplished through the definition of a measure on the data space that takes into account directional dependences of errors, and the use of polar descriptors for straight lines. This strategy improves the robustness by avoiding singularities and non-describable lines.

The methodology is powerful enough to deal with non-normal bivariate heteroscedastic data error models, but can also supersede classical regression methods by making some particular assumptions. An implementation of the methodology for the normal bivariate case is developed and evaluated.

Keywords Errors in both axes; Heteroscedastic data; Linear regression.

Mathematics Subject Classification 62J02.

1. Introduction

Fitting data to straight line models (Rawlings et al., 1998) is one of the most frequently applied statistical procedures. It is widely used in the calibration process in analytical chemistry, in the accelerated lifetime models (González et al., 2009; Vázquez et al., 2007; Yu et al., 2008), and in many other applications where it is required to infer linear trends from a set of experimental data.

Least Squares (LS) techniques (Sayago et al., 2004) for estimating regressions are normally preferred against Bayesian methods because they are easier to interpret and less computationally expensive

The complexity and power of LS methods depends on the error data model. The simplest LS methods, Ordinary Least Squares (OLS) or Weighted Least-Squares (WLS) (Asuero and González, 2007; Mandel and McCrackin, 1988), that consider non null error variance only in the response variable (Y-axis), are analytically resoluble. However, these methods are of limited scope because they assume that the explanatory variable (X-axis) is free of error.

Received June 17, 2009; Accepted October 27, 2009

Address correspondence to Manuel Vázquez, Electronica Fisica, Carretera de Valencia km. 7, Madrid 28031, Spain; E-mail: vazquez@eurtt.upm.es

Other LS methods that take into account errors on any direction as Total Least Squares (TLS) (Markovsky and Van Huffel, 2007), or Bivariate Least Squares (BLS) that is also called Generalized Least Squares (GLS) (Cheng and Riu, 2006), have been also proposed. TLS pursue the minimization of the Euclidean Distance between straight line and data. BLS weights Y-axis errors with factors that take into account variances of data on both axes. BLS has no analytical solution for heteroscedastic data and must therefore be solved by means of numeric algorithms (Martínez et al., 1999).

However, all the above methods only deal with numerical variances, so they are not able to take advantage of more detailed information embedded in the functional knowledge of the data statistics model.

Moreover, they describe straight lines by means of the usual "slope, y-intercept" (m, b) descriptors; therefore, they have a singularity-ambiguity problem in the description of lines that are parallel to Y-axis. So, when they are applied to cases whose solution is close to such lines, the results can be affected by severe and unpredictable inaccuracy.

In this article, we present a methodology that goes around these drawbacks, and makes it possible to deal, from a unified perspective, with the most complex situations where the data uncertainty is modelled by arbitrary statistics. We have developed this methodology, contributing innovations in three stages:

- *The approach.* The Target Radial Deviation Normalized Distance (TRDND) concept, which will be explained in Sec. 2.3, is the fundamental tool that will make it possible to reduce a regression problem statistically formulated to the minimization of a cost function.
  This approach has an essential advantage: it is insensitive to the particular selection of axes on which the data are given. This is a consequence of the fact that TRDND only depends on the statistical data error model and not on the projection of the variances along the particular axes we are working with. In some sense, TRDND let us to introduce in the data space a measure that comes directly from the statistical definition of each datum.
  Moreover, TRDND allows an intuitive conceptual understanding and is powerful enough to be easily generalized to higher dimensions, to deal with non normal errors, and to incorporate information about cross-correlation between the coordinates of each datum.
- *The formulation.* Straight lines are managed through their polar descriptors (D, ψ), instead of the usual "slope, y-intercept" (m, b) descriptors as will be explained in Sec. 2.1. This allows us, to be able to describe straight lines in a unified way (even if they are parallel to axes), and to increase the method robustness by avoiding singularities.
- *The resolution.* As the case of bivariate normal statistics data model is especially useful, a detailed algorithm for this case has been developed. The application of the two former steps to this statistic leads to a system of two high-degree polynomial equations that has not algebraic treatment. So an iterative method, PDIM (Polar Descriptors Iterative Method), for solving it will be presented in Sec. 4. PDIM offers a good trade-off between robustness and computation time, even when it has to process large amounts of data.

The proposed methodology has an interesting feature: when it is customized for normal data statistics, the classical methods mentioned above (OLS, WLS, TLS, and BLS) can be obtained, by making some simple additional assumptions about variances that are described in Table 1. So, the methodology can be viewed as a general and robust framework that includes, as particular cases, all the above-mentioned methods, and can also be applied to solve in a unified way a wider set of problems including multivariate regression (Wilcox, 2009).

As the algebraic expressions involved in PDIM development are quite complex, an algorithm validation has been done by comparing PDIM results with the results of two reference methods. The selected reference methods are: WLS (that considers null variances on X axis), and the reciprocal of WLS (the method obtained by interchanging X and Y axis roles in WLS). Both have been selected with the criterion of being analytically solvable and, at the same time, able to take into account heteroscedasticity.

Additionally, in Sec. 5, some test cases were performed in order to highlight PDIM main features:

(a) The axial independence. The influence of the angle between the regression line and the coordinate axes in the regression quality is compared with the reference methods in Sec. 5.2.1
(b) The precision improvement, with respect to the reference methods, as the heteroscedasticity degree increases (Sec. 5.2.2).
(c) The precision improvement, with respect to the reference methods, as the data inter-axial correlation grows (Sec. 5.2.3).

In order to carry out the tests, a Monte Carlo data synthesis procedure has been designed with the goal of having little directional bias, in the sense that the data can be generated independently of axes orientation. This procedure includes the definition of suitable measures of the degree of heteroscedasticity that are supplied as input parameters to the data synthesis. Also, logarithm measures of the polar descriptors population dispersion are established in order to evaluate the quality of

Table 1
Correspondence between the proposed methodology and the classical methods

Target classical method	Data statistic error model	Variance structure assumptions	Analytical solvability
OLS (Y on X)	Normal	Null X axis variances Equal Y axis variances for all the data	Yes
Reciprocal OLS (X on Y)	Normal	Null Y axis variances Equal X axis variances for all the data	Yes
WLS (HYsY)	Normal	Null X axis variances	Yes
Reciprocal WLS (HYsX)	Normal	Null Y axis variances	Yes
BLS	Normal	None	No
TLS	Normal	Equality of X variance and Y variance for each datum Null inter-axial correlation coefficient for all the data	Yes

results. This scheme is described with detail in Sec. 5, so the methodology behavior can be evaluated and contrasted under any data profile.

2. Terminology and Tools

A first assumption that is commonly made in regression problems is the statistical independence of any datum with respect to the others. Although this is not ever rigorously true in all situations, it allows each datum to be considered as an individual entity that can be modeled by a single probability function. When this assumption is satisfied, we will say that each datum is an *uncertain point* in the data space and the data set constitutes a *cloud* of uncertain points. An uncertain point can be fully defined given its measured values $(\tilde{x}_i, \tilde{y}_i)$ on an (X, Y) axis reference frame, and the error probability density function: $f_{\varepsilon i}(\varepsilon x_i, \varepsilon y_i)$, that is a function that depends only on error components εx_i, εy_i along the (X, Y) axes. So the true point value (x_i, y_i) can be obtained by means of (1):

$$\begin{aligned} x_i &\equiv \tilde{x}_i - \varepsilon x_i \\ y_i &\equiv \tilde{y}_i - \varepsilon y_i \end{aligned} \tag{1}$$

Once $f_{\varepsilon i}(\varepsilon x_i, \varepsilon y_i)$ is known, it is obvious to evaluate $f_i(x_i, y_i)$, the ith datum probability density function, by means of $f_i(x_i, y_i) = f_{\varepsilon i}(\tilde{x}_i - x_i, \tilde{y}_i - y_i)$.

$f_i(x_i, y_i)$ is usually displayed by means of isodensity sets (Fig. 1). In non degenerate cases, isodensity sets are lines (isodensity lines) that link data space points that have the same probability density.

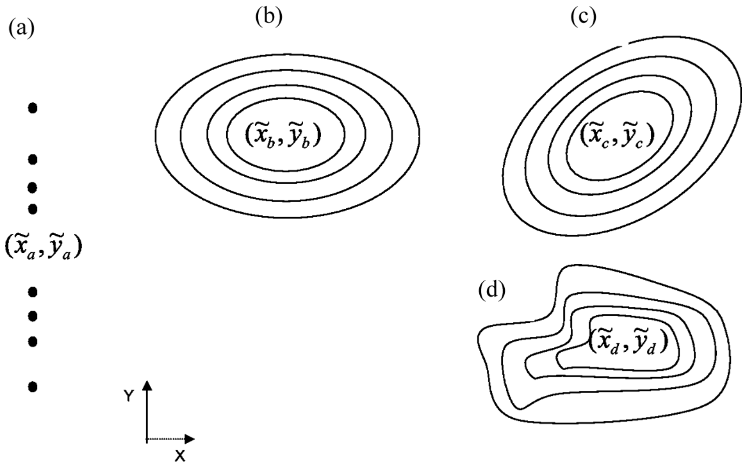
(a)
(b)
(c)
$(\tilde{x}_a,\tilde{y}_a)$
$(\tilde{x}_b,\tilde{y}_b)$
$(\tilde{x}_c,\tilde{y}_c)$
(d)
$(\tilde{x}_d,\tilde{y}_d)$
Y
X

Figure 1. Isodensity sets for uncertain points: (a) a normal point with null deviation in X error component (degenerate Y univariate case); (b) a normal point with non null but uncorrelated deviation in both error components (Bivariate inter-axially independent case); (c) a normal point with non zero deviation in both variables and non zero cross-correlation (bivariate inter-axially non-independent case); (d) a non normal uncertain point.

We will say that an uncertain point is *normal* if the pair $\varepsilon x_i, \varepsilon y_i$ follow a zero mean normal bivariate statistic (2):

$$f_{\varepsilon i}(\varepsilon x_i, \varepsilon y_i) = \frac{1}{2\pi\sigma_{x_i}\sigma_{y_i}\sqrt{1-\left(\rho_{xy_i}\right)^2}} \times \exp\left[-\frac{1}{2}\frac{1}{1-\left(\rho_{xy_i}\right)^2}\left(\frac{\varepsilon x_i^2}{\sigma_{x_i}^2}+\frac{\varepsilon y_i^2}{\sigma_{y_i}^2}-2\frac{\rho_{xy_i}\varepsilon x_i \varepsilon y_i}{\sigma_{x_i}\sigma_{y_i}}\right)\right]. \tag{2}$$

In (2), σ_{xi}, σ_{yi} are the marginal deviations of the two error components along axes, and ρ_{xyi} is the inter-axial correlation coefficient. Once given these three parameters, the inter-axial cross-covariance σ^2_{xyi} can be calculated as $\sigma_{xyi} = \rho_{xyi}\sigma_{xi}\sigma_{yi}$. If the error components are independent, both ρ_{xyi} and σ_{xyi} will be zero.

We will say that a cloud is a *normal cloud* when all its uncertain points are normal, so a normal cloud can be completely described by a data structure that has five numeric descriptors $\{\sigma_{xi}, \sigma_{yi}, \rho_{xyi}, \tilde{x}_i, \tilde{y}_i\}$ for each of its uncertain points.

Isodensity lines for normal points are the locus where the probability density function (2) is constant. So, as a function of the error coordinates $(\varepsilon x_i, \varepsilon y_i)$, isodensity lines will have equations as (3):

$$\frac{\varepsilon x_i^2}{\sigma_{x_i}^2}+\frac{\varepsilon y_i^2}{\sigma_{y_i}^2}-2\frac{\sigma_{xy_i}\varepsilon x_i \varepsilon y_i}{\sigma_{x_i}^2\sigma_{y_i}^2} = K^2, \tag{3}$$

where K is a parameter that labels each isodensity line. By (3), it is clear that isodensity sets of normal uncertain points are ellipses centered on the origin in the error space (or ellipses centered on the measurement $(\tilde{x}_i, \tilde{y}_i)$ in the data space).

2.1. *Straight Line Polar Descriptors*

Using usual "slope, y-intercept" descriptors (m, b) or "counter-slope, x-intercept" (n, a) to represent straight lines in the forms (4) or (5) leads to unsafe computations since both parameters can be unbounded:

$$y = mx + b \tag{4}$$

$$x = ny + a. \tag{5}$$

That occurs because as lines get more strongly sloped, the magnitudes of m and b grow towards infinity. So for computational purposes, it is better to define straight lines with their polar descriptors (D, ψ). Polar descriptors of a straight line should not be confused with the polar coordinates of a point.

Straight line polar descriptors meaning can be seen in Fig. 2. Here, D is the distance from the origin to the straight line, and Ψ is the angle that defines the closest point (Nearby Point) of the straight line to the origin O. So, the (D, ψ) polar descriptors define straight lines in a smooth and bijective way. The domains for both descriptors are: $D \geq 0$ and $\psi \in [0, 2\pi)$.

The Cartesian coordinates (x, y) of any point R of the (D, ψ) straight line r should satisfy Eq. (6):

$$r = \{R \equiv (x, y)/x\cos(\Psi) + y\sin(\Psi) - D = 0\}. \tag{6}$$

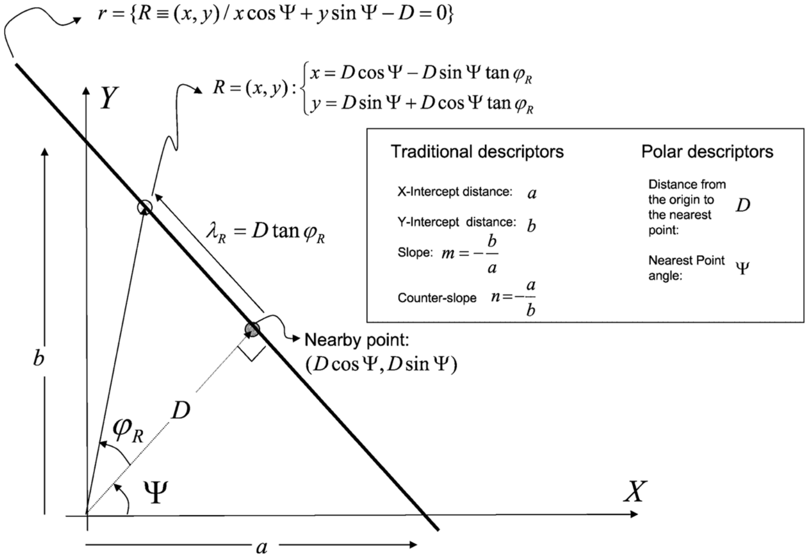
$r = \{R \equiv (x,y) / x\cos\Psi + y\sin\Psi - D = 0\}$
$R = (x,y): \begin{cases} x = D\cos\Psi - D\sin\Psi\tan\varphi_R \\ y = D\sin\Psi + D\cos\Psi\tan\varphi_R \end{cases}$
Y
$\lambda_R = D\tan\varphi_R$
Traditional descriptors
X-Intercept distance: a
Y-Intercept distance: b
Slope: $m = -\frac{b}{a}$
Counter-slope $n = -\frac{a}{b}$
Polar descriptors
Distance from the origin to the nearest point: D
Nearest Point angle: Ψ
Nearby point:
$(D\cos\Psi, D\sin\Psi)$
b
D
φ_R
Ψ
X
a

Figure 2. Straight line polar descriptors.

So, its coordinates can be calculated, once given the deflection angle φ_R of point R as (7):

$$R \equiv \begin{cases} x = D\cos(\Psi) - D\sin(\Psi)\tan(\varphi_R) \\ y = D\sin(\Psi) + D\cos(\Psi)\tan(\varphi_R). \end{cases} \tag{7}$$

Or as a function of the alternative parameter $\lambda_R = D\tan\varphi_R$ that describes a point R by its signed distance λ_R from the nearby point (8):

$$\begin{aligned} x &= D\cos\psi - \lambda_R\sin\psi \\ y &= D\sin\psi + \lambda_R\cos\psi \end{aligned} \tag{8}$$

2.2. *Radial Deviation*

Radial deviation is a measure that quantifies the error an uncertain point has along a particular direction. The radial deviation can be calculated as a function of the angle θ that defines a given direction. Expressing the error components of the uncertain point i in polar coordinates we have (9):

$$\begin{aligned} \varepsilon x_i &= r\cos\theta \\ \varepsilon y_i &= r\sin\theta \end{aligned} \tag{9}$$

The radial deviation for a given point $\sigma_{ri}(\theta)$ is defined (see Fig. 3) as the standard deviation of the distance, conditioned to be in the semi-straight line that

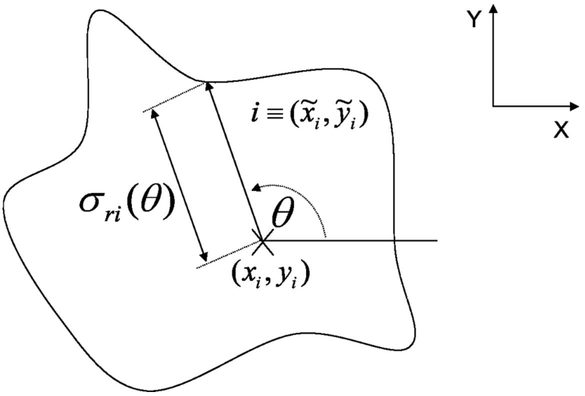

Figure 3. Experimental points are modeled by uncertain points centered in the measured values. The Radial Deviation contour is defined by the Radial Deviation of the error for different θ values.

starts in $(\tilde{x}_i, \tilde{y}_i)$ and forms an angle θ with the X-axis, and can be calculated as (10):

$$\sigma^2_{ri}(\theta) = \frac{\int_0^\infty r^2 f_{\varepsilon i}(r\cos\theta, r\sin\theta)dr}{\int_0^\infty f_{\varepsilon i}(r\cos\theta, r\sin\theta)dr}, \tag{10}$$

where $f_{\varepsilon i}(r\cos\theta, r\sin\theta)$ is the error probability density function in a specific direction defined by the angle θ. In Fig. 3 , the locus defined by the radial deviation for different θ values has been represented.

Radial deviation, $\sigma_{ri}(\theta)$, should not be mistaken with $\sigma_{Di}(\theta)$: the marginal directional deviation (Duda et al., 1997), that can be calculated by means of (11):

$$\sigma^2_{Di}(\theta) = 2\int_0^\infty \int_{-\infty}^\infty u^2 f_{\varepsilon i}(u\cos\theta - v\sin\theta, u\sin\theta + v\cos\theta)dv\,du \tag{11}$$

2.2.1. *Radial Deviation of a Normal Point.* It can be easily demonstrated, by transforming (2) by (9), that the error associated with a normal point along a particular radius (defined by the angle θ), follows a normal distribution. Moreover, the radial deviations contour $\sigma_{ri}(\theta)$ (that is the locus described by a vector that forms an angle θ with the X-axis, and whose modulus is precisely $\sigma_{ri}(\theta)$), draws an ellipse that matches with an isodensity line whose equation is (12):

$$\sigma^2_{ri}(\theta) = \frac{\sigma^2_{xi}\sigma^2_{yi}(1-\rho^2_{xyi})}{\sigma^2_{xi}\sin^2\theta - 2\rho_{xyi}\sigma_{xi}\sigma_{yi}\sin\theta\cos\theta + \sigma^2_{yi}\cos^2\theta} \tag{12}$$

On the other hand, nevertheless, the error along a defined direction θ also follows a normal distribution, the function that determines the marginal directional deviation $\sigma_{Di}(\theta)$ is, even for normal uncertain points, very complex and gives a non-ellipse line as is depicted in Fig. 4. Radial deviation and marginal directional deviation only matches on the principal axes (given by the eigenvectors of the covariance matrix) where they take the values σ_{MINi} and σ_{MAXi}. Moreover, marginal directional deviation

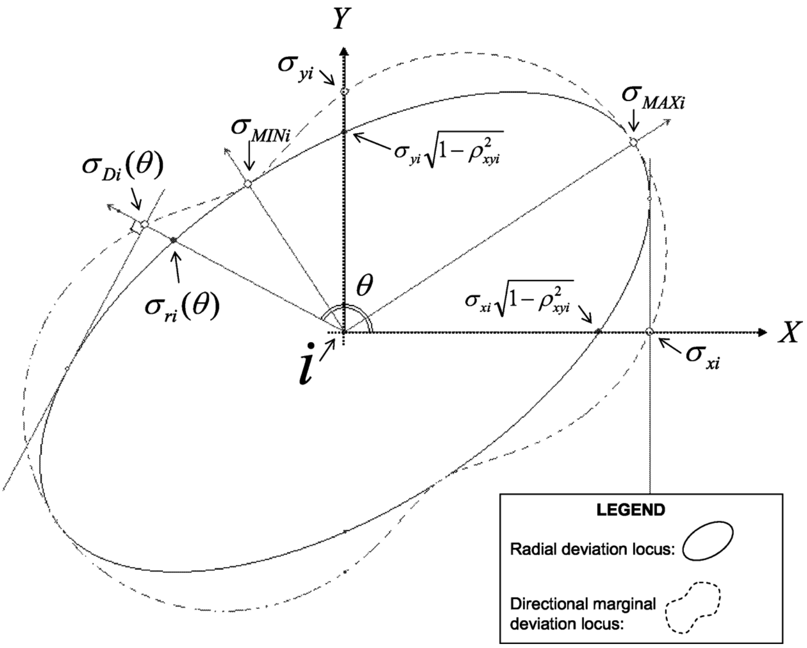

Figure 4. Radial deviation (solid) and marginal directional deviation (dashed) angular loci for an uncertain normal point.

along the X and Y axes match the density (σ_{xi}, σ_{yi}) descriptors.

$$\begin{aligned} \sigma_{Di}(0) &= \sigma_{xi} \\ \sigma_{Di}\left(\frac{\pi}{2}\right) &= \sigma_{yi}. \end{aligned} \tag{13}$$

However, radial deviations along descriptive axes (14) do not match density descriptors when there are some inter-axial correlation $\rho_{xyi} \neq 0$.

$$\begin{aligned} \sigma_{ri}(\theta = 0) &= \sigma_{xi}\sqrt{1 - \rho_{xyi}^2} \\ \sigma_{ri}\left(\theta = \frac{\pi}{2}\right) &= \sigma_{yi}\sqrt{1 - \rho_{xyi}^2} \end{aligned} \tag{14}$$

2.3. *Target Radial Deviation Normalized Distance (TRDND)*

The proposed linear regression methodology is a weighted method. Data with lower uncertainty will have higher weight in the cost function. The uncertainty will be taken into consideration by using the radial deviation concept explained in the previous section. This leads to the definition of a dimensionless measure that will be called "Target Radial Deviation Normalized Distance" (TRDND).

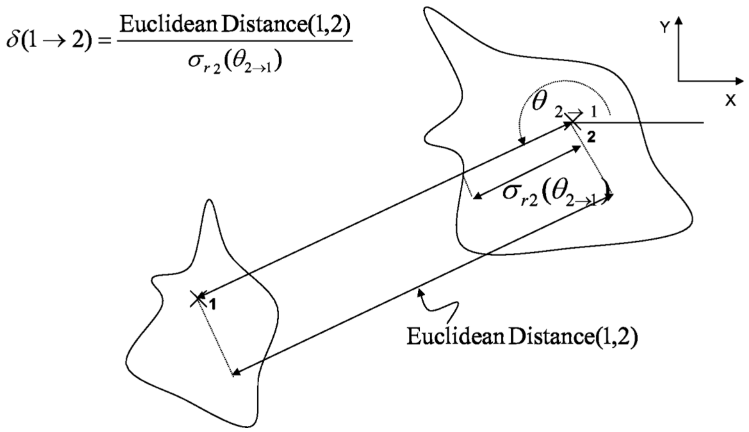
$\delta(1 \rightarrow 2) = \frac{\text{Euclidean Distance(1,2)}}{\sigma_{r2}(\theta_{2\rightarrow 1})}$
Y
X
$\theta_{2\rightarrow 1}$
2
$\sigma_{r2}(\theta_{2\rightarrow 1})$
1
Euclidean Distance(1,2)

Figure 5. TRDND from a point **1** to an uncertain point **2**.

2.3.1. *TRDND from one Point to Another Point.* TRDND is a "distance-like" measure between two points **1** and **2**. One of them, say **1**, is considered the origin and the other, **2**, the target, so we will speak about TRDND *from* **1** *to* **2.** TRDND from **1** to **2** will be denoted as $\delta(1 \rightarrow 2)$. TRDND only takes into account the uncertainty of the target point so TRDND is not, in general, a symmetric measure. This means that $\delta(1 \rightarrow 2) = \delta(2 \rightarrow 1)$ is not necessarily a true assertion.

The idea behind TRDND is to normalize the conventional Euclidean distance between the centers of both points by the radial deviation of the target point along the direction that goes towards to the origin point: $\sigma_{r2}(\theta_{2\rightarrow 1})$. Here, $\theta_{2\rightarrow 1}$ is the angle between the X axis and the semi-straight line that comes from **2** and goes to **1**. So TRDND results in a dimensionless measure given by (15):

$$\delta^2(1 \rightarrow 2) = \frac{(\tilde{x}_2 - \tilde{x}_1)^2 + (\tilde{y}_2 - \tilde{y}_1)^2}{\sigma_{r2}^2(\theta_{2\rightarrow 1})}, \tag{15}$$

where $\sigma_{r2}^2(\theta_{2\rightarrow 1})$ can be calculated by means of Eq. (10). It's worth emphasizing that TRDND only depends on the error model of the target point.

2.3.1.1. *TRDND from a Point to a normal point.* As radial deviation for normal points along any arbitrary direction can be calculated by (12), it is easy to evaluate the TRNDN for normal targets as (16):

$$\delta^2(1 \rightarrow 2) = \frac{(\tilde{x}_2 - \tilde{x}_1)^2 \sigma_y^2 + (\tilde{y}_2 - \tilde{y}_1)^2 \sigma_x^2 - 2(\tilde{y}_2 - \tilde{y}_1)(\tilde{x}_2 - \tilde{x}_1)\sigma_{xy}}{\sigma_x^2\sigma_y^2 - \sigma_{xy}^2}. \tag{16}$$

When the target is a normal point, TRDND can also be understood from the Mahalanobis concept (Mahalanobis, 1936); that is, $\delta(1 \rightarrow 2)$ matches the Mahalanobis distance between both points when the covariance matrix is imposed by the target point **2**.

2.3.2. *TRDND from a Line to a Point.* In the same way that occurs with Euclidean distances, it is possible to extend TRDND definition to embrace distances from a line to a target point. We will consider TRDND from a line r, to a target point, i as the minimum TRDND from all the points $R \in r$ to the point i (17):

$$\delta_i^2\left[r \to i\right] = Minimum\left[\delta^2\left[R \to i\right]\ \forall R \in r\right]. \tag{17}$$

According to (17), we can obtain an explicit equation for TRNDN from a straight line when the target is a normal point. By replacing straight line polar descriptors D, Ψ in the TRDND Eq. (16) for the different points of the straight line (defined in terms of different values of λ_R) and the datum i defined as a normal uncertain point we obtain (18):

$$\begin{aligned}\delta^2\left[R \to i\right] = \frac{1}{1-\rho_{xy_i}^2}\Bigg(&\frac{(D\cos(\Psi) - \tilde{x}_i - \sin(\Psi)\lambda_R)^2}{\sigma_x^2} \\ &+ \frac{(D\sin(\Psi) - \tilde{y}_i + \cos(\Psi)\lambda_R)^2}{\sigma_y^2} \\ &- \frac{2\,(D\sin(\Psi) - \tilde{y}_i + \cos(\Psi)\lambda_R)\,(D\cos(\Psi) - \tilde{x}_i + \sin(\Psi)\lambda_R)\,\rho_{xy}}{\sigma_x\sigma_y}\Bigg)\end{aligned} \tag{18}$$

Constraining the derivative with respect to λ_R of this function to be zero, we can evaluate λ_{RMIN}, the value that identifies the point of the straight line that minimizes the TRDND. Eventually, by replacing the obtained value of λ_{RMIN} in (18) we obtain the TRDND from the straight line to a normal point as (19):

$$\begin{aligned}\delta^2\left[r \to i\right] &= \frac{(-D + \cos(\Psi)\tilde{x}_i + \sin(\Psi)\tilde{y}_i)^2}{\sigma_{x_i}^2\cos^2(\Psi) + \rho_{xy_i}\sigma_{x_i}\sigma_{y_i}\sin(2\Psi) + \sigma_{y_i}^2\sin^2(\Psi)} \\ &= \frac{(-D + \cos(\Psi)\tilde{x}_i + \sin(\Psi)\tilde{y}_i)^2}{\sigma_i^2}.\end{aligned} \tag{19}$$

Equation (19) allows interpreting $\delta\left[r \to i\right]$ as the Euclidean distance between the line and the point, divided by a partition value: σ_i (20) that does not depend on D:

$$\sigma_i^2 = \sigma_{x_i}^2\cos^2(\Psi) + \rho_{xy_i}\sigma_{x_i}\sigma_{y_i}\sin(2\Psi) + \sigma_{y_i}^2\sin^2(\Psi). \tag{20}$$

It's worth mentioning that σ_i does not match, in general, with marginal directional deviation (11) nor radial deviation (12).

2.3.3. *TRDND from One Line to a Cloud of Points.* Following the same strategy, we can define the squared TRDND from a line: r, to a cloud: $I = \{1, 2 \ldots i \ldots n\}$: $\delta^2\left[r \to I\right]$ to be equal to the variance residual (21):

$$\delta^2\left[r \to I\right] = \frac{1}{n-2}\sum_{i \in I}\delta^2(r \to i). \tag{21}$$

If we call Q_i^2 to the squared TRDND from the straight line to the datum i, we have:

$$Q_i^2 = \delta^2(r \to i). \tag{22}$$

By simply applying this definition to (21), we have (23):

$$\delta^2\left[r \to I\right] = \frac{1}{n-2}\sum_{i\in I} Q_i^2. \tag{23}$$

$\delta\left[r \to I\right]$ can be understood as a measure of the dissimilarity between a line r and a cloud I, that takes into account the uncertainties of the cloud's points in the direction that points to r. In particular, if the points of I lie on r, then $\delta\left[r \to I\right] = 0$.

For straight lines and normal clouds we can use (19) to obtain (24):

$$Q_i^2 = \frac{(-D + \cos(\Psi)\tilde{x}_i + \sin(\Psi)\tilde{y}_i)^2}{\sigma_i^2}. \tag{24}$$

3. TRDND Formulation of Regression Problems

Given a cloud I, made up of n points $I = \{1, 2, 3, \ldots i \ldots n\}$, the proposed methodology reduces the regression problem to the problem of finding the line $r_{guess} \equiv (D_{guess}, \Psi_{guess})$ that minimizes TRDND from r_{guess} to I. The couple $(D_{guess}, \Psi_{guess})$ can be evaluated by forcing the derivatives of (23) to zero. Thus, we obtain the constraints (25):

$$\begin{array}{l} \dfrac{\partial\delta^2\left[r_{guess} \to I\right]}{\partial D_{guess}} = 0 \\ \dfrac{\partial\delta^2\left[r_{guess} \to I\right]}{\partial \psi_{guess}} = 0 \end{array}. \tag{25}$$

As there is one constraint for each unknown, the system has a well-defined straight line solution.

As it has been explained above, this methodology constitutes a generalized framework that can be particularized to reproduce classical straight line regression results. Typically, these particularizations are the normality of the clouds, besides others common sense assumptions about means and variances. For example, we can obtain BLS cost function by using our methodology under the normal cloud hypothesis, by using Eq. (19), and reformulating it by substituting the polar descriptors (D, Ψ) by the classical slope-y intercept descriptors (m, b) with the help of the transformations (26):

$$\begin{array}{l} D = b\sin(\Psi) \\ m = -ctg(\Psi) \end{array}. \tag{26}$$

That leads to Eq. (27):

$$\delta^2\left[r \to i\right] = \frac{(\tilde{y}_i - m\tilde{x}_i - b)^2}{\sigma_{yi}^2 - 2b\sigma_{xyi} + b^2\sigma_{xi}^2}. \tag{27}$$

(27) is formally identical to BLS formula (7) of Sayago et al. (2004), so our methodology gives the same results than BLS if we assume:

(a) The normality of the cloud.

(b) The identification between the squared standard deviations of marginal error densities ($\sigma_{xi}^2, \sigma_{yi}^2$) and the variances ($s_{xi}^2, s_{yi}^2$) supplied to BLS.

(c) The identification between the expected values of data densities and the measured values supplied to BLS.

Other classical methods can be obtained in a similar way. Table 1 shows the assumptions that must be done in order to force the results of the proposed methodology to be the same that are obtained by some classical methods. However, even for these cases, the improved robustness that comes from the use of polar descriptors for straight lines and from the unifications of multiple methods in a single algorithm gives a clear advantage to the methodology:

4. PDIM: An Iterative Method for Solving the Regression Problem

In this section, a numerical algorithm: PDIM (Polar Descriptors Iterative Method) will be developed for solving straight line fitting regression problems based in the proposed methodology for normal clouds. According to the former section, our objective is to find the straight line $r_{guess} \equiv (D_{guess}, \Psi_{guess})$ that minimizes $\delta^2\left[r_{guess} \to I\right]$ (23). As $\frac{1}{n-2}$ is a constant factor for a given problem, it will be enough to minimize the cost function Q^2 defined by (28):

$$Q^2 = \sum_{i \in I} Q_i^2. \tag{28}$$

Under the assumption of normal clouds, the partition values σ_i (20) are all polynomials of degree two in $z = \cos\Psi$ that can have different factorization for each datum. So the sum Q^2 becomes an algebraic fraction whose numerator's degree can grow up to a value of $2n$, making the algebraic resolution intractable even for low n values. PDIM must be therefore an approximate method.

The main idea behind PDIM is to determine a succession of k straight lines $r[k] \equiv (D[k], \Psi[k])$ that goes in the proximity of a minimum of (28), $(D_{guess}, \Psi_{guess})$, as the succession index k grows. To do that, in each step k, we will find the values $(D[k], \Psi[k])$ that minimize (29):

$$Q^2[k] = \sum_{i \in I} Q_i^2[k] = \sum_{i \in I} \frac{(\tilde{x}_i \cos(\Psi[k]) + \tilde{y}_i \sin(\Psi[k]) - D_i[k])^2}{\sigma_i^2[k-1]}. \tag{29}$$

Equation (29) assumes that the actual partition values $\sigma_i[k]$ can be approximated by its previous values: $\sigma_i[k-1] \approx \sigma_i[k]$. Moreover, $\sigma_i[k-1]$ depends only on the angle descriptor of the previous regression line: $\Psi[k-1]$, and can be calculated at the beginning of the kth step by means of (30):

$$\begin{aligned}\sigma_i^2[k-1] = {} & \sigma_{x_i}^2 \cos^2(\Psi[k-1]) + \rho_{xy_i}\sigma_{x_i}\sigma_{y_i} \sin(2\Psi[k-1]) \\ & + \sigma_{y_i}^2 \sin^2(\Psi[k-1]).\end{aligned} \tag{30}$$

This assumption maintains all the denominators of (29) independent of the actual step descriptors, $(D[k], \Psi[k])$, so we can add all its the terms in a homogeneous way in order to obtain an algebraically solution for $(D[k], \Psi[k])$.

The validity of (29) can be justified by the following fact: if we derive (29) with respect to ($D[k]$, $\Psi[k]$), and constrain them to be zero, we get to a set of two equations that let us to determine ($D[k]$, $\Psi[k]$) as a function of ($D[k-1]$, $\Psi[k-1]$). It is clear that if the succession $r[k]$ stabilizes, then $\Psi[k] \cong \Psi[k-1]$ and $D[k] \cong D[k-1]$. In particular, by (30), this means that the condition $\sigma_i[k-1] \cong \sigma_i[k]$ is fulfilled and (29) is valid. Therefore, the convergence of ($D[k]$, $\Psi[k]$) is a criterion for both: the stationary character of the limit, and the validity of (29).

By making the derivative of (29) with respect to $D[k]$, we get a first constrain (31):

$$\begin{aligned} 0 &= -\frac{1}{2}\frac{\partial}{\partial D[k]} Q_i^2[k] = \sum_{i\in I} -\frac{1}{2}\frac{\partial}{\partial D[k]} Q_i^2[k] \\ &= \sum_{i\in I} \frac{\tilde{x}_i \cos(\Psi[k]) + \tilde{y}_i \sin(\Psi[k]) - D[k]}{\sigma_i^2[k-1]}. \end{aligned} \tag{31}$$

Now, in order to simplify the equations we define the following momenta (32):

$$\begin{aligned} \gamma_i[k] &= \frac{1}{\sigma_i[k]} \\ \alpha_{gg}[k] &= \sum_{i\in I} \gamma_i^2[k] \\ \alpha_{xgg}[k] &= \sum_{i\in I} \tilde{x}_i \gamma_i^2[k] \\ \alpha_{ygg}[k] &= \sum_{i\in I} \tilde{y}_i \gamma_i^2[k] \\ \alpha_{xxgg}[k] &= \sum_{i\in I} \tilde{x}_i^2 \gamma_i^2[k] \\ \alpha_{xygg}[k] &= \sum_{i\in I} \tilde{x}_i \tilde{y}_i \gamma_i^2[k] \\ \alpha_{yygg}[k] &= \sum_{i\in I} \tilde{y}_i^2 \gamma_i^2[k]. \end{aligned} \tag{32}$$

These conventions let us add (31) and express it as:

$$0 = \alpha_{xgg}[k-1]\cos(\Psi[k]) + \alpha_{ygg}[k-1]\sin(\Psi[k]) - \alpha_{gg}[k-1]D[k]. \tag{33}$$

Isolating $D[k]$ in (31), we have:

$$D[k] = \frac{\alpha_{xgg}[k-1]\cos(\Psi[k]) + \alpha_{ygg}[k-1]\sin(\Psi[k])}{\alpha_{gg}[k-1]}. \tag{34}$$

Now, we build a second constrain by forcing the derivative of (29), now with respect to $\Psi[k]$, to be zero.

$$0 = \frac{1}{2}\frac{\partial}{\partial \Psi[k]} Q^2[k] = \sum_{i\in I} \frac{1}{2}\frac{\partial}{\partial \Psi[k]} Q_i^2[k]. \tag{35}$$

That yields:

$$0 = \sum_{i \in I} \frac{(-\tilde{x}_i \sin(\Psi[k]) x_i + \tilde{y}_i \cos(\Psi[k]))\,(\tilde{x}_i \cos(\Psi[k]) + \tilde{y}_i \sin(\Psi[k]) - D[k])}{\sigma_i^2[k-1]}. \tag{36}$$

Substituting the value of $D[k]$ given by (34) into (36), grouping the terms with the same dependence on $\Psi[k]$ and using the momenta defined in (32), we come, after some algebra stuff, to (37):

$$\begin{aligned} 0 = {} & \left(\alpha_{xygg}[k-1] - \frac{\alpha_{xgg}[k-1]\alpha_{ygg}[k-1]}{\alpha_{gg}[k-1]}\right)\cos^2(\Psi[k]) \\ & - \left(\alpha_{xygg}[k-1] - \frac{\alpha_{xgg}[k-1]\alpha_{ygg}[k-1]}{\alpha_{gg}[k-1]}\right)\sin^2(\Psi[k]) \\ & + \left(\alpha_{yygg}[k-1] - \alpha_{xxgg}[k-1] + \frac{\alpha_{xgg}^2[k-1] - \alpha_{ygg}^2[k-1]}{\alpha_{gg}[k-1]}\right)\sin(\Psi[k])\cos(\Psi[k]). \end{aligned} \tag{37}$$

And now by calling

$$\begin{aligned} A[k-1] &= \alpha_{xygg}[k-1] - \frac{\alpha_{xgg}[k-1]\alpha_{ygg}[k-1]}{\alpha_{gg}[k-1]} \\ B[k-1] &= \frac{1}{2}\left(\alpha_{yygg}[k-1] - \alpha_{xxgg}[k-1] + \frac{\alpha_{xgg}^2[k-1] - \alpha_{ygg}^2[k-1]}{\alpha_{gg}[k-1]}\right), \end{aligned} \tag{38}$$

we can simplify (37) by using (38) in the form (39):

$$0 = A[k-1]\left(\cos^2(\Psi[k]) - \sin^2(\Psi[k])\right) + 2B[k-1]\sin(\Psi[k])\cos(\Psi[k]). \tag{39}$$

Now, by using the trigonometric formulae for double angle, we can write (39) in an even more synthetic way.

$$0 = A[k-1]\cos(2\Psi[k]) + B[k-1]\sin(2\Psi[k]). \tag{40}$$

So the system (31), (35), has an analytic solution given by (41) and (42):

$$\Psi[k] = \frac{1}{2} Arc\tan\left(\frac{-B[k-1]}{A[k-1]}\right) + p\frac{\pi}{2} \quad (\text{with } p \in \mathbf{Z}) \tag{41}$$

$$D[k] = \frac{\alpha_{xgg}[k-1]\cos(\Psi[k]) + \alpha_{ygg}[k-1]\sin(\Psi[k])}{\alpha_{gg}[k-1]}. \tag{42}$$

The solution for $\Psi[k]$: (41), has four discernible branches corresponding to the values of $p \in \{0, 1, 2, 3\}$. The last task in each step should be to select which of them actually corresponds to the minimum of $Q[k]$.

Each one of the $\Psi[k]$ solutions leads to a corresponding $D[k]$ value by Eq. (42). From among these four solutions, two of them give negative values for $D[k]$, so they are nonsense. Focusing on the other two, one of them corresponds to a maximum, and the other to a minimum. In order to select the minimum, we have to compare the cost function value for both solutions, and select the solution that gives the cost function $Q[k]$ whose value is the smallest. So, in order to perform the selection, it is required to obtain $Q[k]$ as a function of the momenta (32). We can do it by expanding (29) in a sum of terms, and doing the sum of the series by using the momenta (32). That yields (43):

$$\begin{aligned} Q^2[k] = {} & \alpha_{xxgg}[k]\cos^2(\Psi[k]) + 2\alpha_{xygg}[k]\cos(\Psi[k])\sin(\Psi[k]) \\ & + \alpha_{yygg}[k]\cos^2(\Psi[k]) - 2\alpha_{xgg}[k]D[k]\cos(\Psi[k]) \\ & - 2\alpha_{ygg}[k]D[k]\sin(\Psi[k]) + \alpha_{gg}[k]D^2[k]. \end{aligned} \tag{43}$$

Equation (43) is also suited for calculating the TRDND from the regression line to the cloud $\delta[r \to I]$, which constitutes a quality measurement of the regression.

Once having Q, TRDND can be calculated as: $\delta[r \to I] = \frac{1}{\sqrt{n-2}}Q$.

It is convenient to point out that this scheme does not assure the succession convergence towards the desired minimum. Depending on the initial value of descriptors, the succession could diverge or be chaotic. Moreover, if the data are ill conditioned and the cost function has multiple minima, it can go towards a local minimum different to the desired absolute minimum.

So, it is required to explore the descriptor's initial value space in order to reject the non convergent successions and select, among the convergent ones, the one that gives place to the least value of the cost function (43). Anyway, this exploration process is greatly simplified by the fact that the succession has been carefully constructed in such a way that its dynamic is fully determined by the initial value of the single Ψ descriptor,$\Psi[0]$, because in each step, the $D[k]$ descriptor comes algebraically determined by the $\Psi[k]$ value through (42). This reduces the exploration to the one-dimensional interval: $\Psi[0] \in [0, 2\pi)$.

5. Evaluation and Results

In this section, we will show the scheme that was developed in order to evaluate linear regression methods, and its application to highlight PDIM features under some quite general conditions.

Monte Carlo techniques are used in order to synthesize trial clouds that follow a desired profile. The objective is to achieve a good (understandable and controllable) parameterization of clouds' behavior.

In order to fulfil this goal, the scheme takes advantage of polar description of linear laws. Uncertain points are constructed by selecting randomly at the beginning their deflection angles φ seen from the origin (Fig. 2), and then using (7) in order to obtain the coordinates of its centers, (instead of starting from choosing at random one single coordinate (x or y), and using (4) or (5) in order to determine the other one, as is done usually). This strategy has two advantages.

1. Any underlying linear law can be modeled, including lines that are parallel to the Y-axis.
2. The cloud scattering profile description gets decoupled from the underlying law. That is: the clustering features of the clouds can be specified without taking into account the slope of the law or the axis system on which the data are given.

Another feature of the scheme is the use of logarithms units (dB) for all positive descriptors and error measures; it has the advantage of widening the range of situations that are described in the graphs, and isolating their interpretations from the units on which data are given.

5.1. *Evaluation Procedure and Cloud Synthesis*

The guidelines of the evaluation procedure are depicted in Fig. 6. A population $I\{t\}$ of T clouds ($1 \leq t \leq T$) is synthesized by Monte Carlo techniques, according with a set of eleven real parameters (cloud profile) that describe statistically the features of the kind of cloud against which one want to test the methodology.

These 11 cloud profile descriptors (that appears in the leftmost box inside the population box of Fig. 6), can be grouped in four groups: from below to above:

(1) The true law descriptors $r_{true} \equiv (D_{true}, \Psi_{true})$ that determines the line where the data centres lie.

(2) The number of points n that make up the clouds, and the flare angle interval $[\varphi_{\min}, \varphi_{\max}]$, that is the range of angles seen from the origin (and measured with respect to the direction of the nearest point of the true law line) from where the data centres can be chosen (see Fig. 2). Thus, the flare angle interval controls directly the asymmetry of the cloud with respect to the nearest point and the amount of dispersion of data centers along the law line.

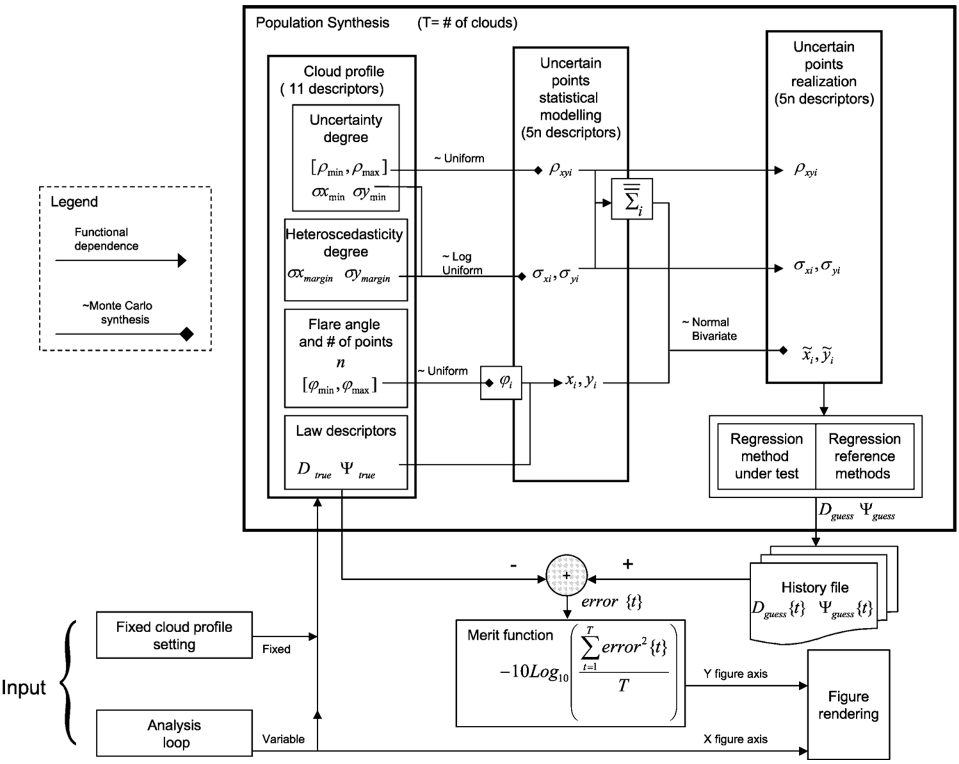
Population Synthesis (T= # of clouds)
Cloud profile (11 descriptors)
Uncertainty degree
$[\rho_{\min}, \rho_{\max}]$
$\sigma x_{\min}$ $\sigma y_{\min}$
Heteroscedasticity degree
σx_{margin} σy_{margin}
Flare angle and # of points
n
$[\varphi_{\min}, \varphi_{\max}]$
Law descriptors
D_{true} Ψ_{true}
~ Uniform
~ Log Uniform
~ Uniform
Uncertain points statistical modelling (5n descriptors)
ρ_{xyi}
σ_{xi}, σ_{yi}
φ_i
x_i, y_i
Σ_i
~ Normal Bivariate
Uncertain points realization (5n descriptors)
ρ_{xyi}
σ_{xi}, σ_{yi}
$\tilde{x}_i, \tilde{y}_i$
Regression method under test
Regression reference methods
D_{guess} Ψ_{guess}
Legend
Functional dependence
~Monte Carlo synthesis
-
+
+
error $\{t\}$
History file
$D_{guess}\{t\}$ $\Psi_{guess}\{t\}$
Merit function
$-10Log_{10}\left(\frac{\sum_{t=1}^{T} error^2\{t\}}{T}\right)$
Y figure axis
X figure axis
Figure rendering
Input
Fixed cloud profile setting
Fixed
Analysis loop
Variable

Figure 6. Test methodology explanatory diagram.

(3) The descriptors of the uncertainty degree of the points of the cloud, namely the allowed interval for inter-axial cross correlation $[\rho_{\min}, \rho_{\max}]$, and the minimum allowed value of the deviation along each data axis: $\sigma x_{\min}(dB)$, and $\sigma y_{\min}(dB)$. The deviations are given in logarithms units (with respect to the unit measure of D_{true}); that is (44):

$$\begin{aligned}\sigma x_{\min}(dB) &= 20\,Log_{10}(\sigma x_{\min})\\ \sigma y_{\min}(dB) &= 20\,Log_{10}(\sigma y_{\min}).\end{aligned} \tag{44}$$

(4) The descriptors of the heteroscedasticity degree. These are the amplitude margin of directional deviation of error on X and Y directions: σx_{margin} and σy_{margin}. These two values are also supplied in dB, so the maximum values $(\sigma x_{\max}, \sigma y_{\max})$ of the directional deviations in dB can be calculated as (45):

$$\begin{aligned}\sigma x_{\max}(dB) &= \sigma x_{\min}(dB) + \sigma x_{margin}(dB)\\ \sigma y_{\max}(dB) &= \sigma y_{\min}(dB) + \sigma y_{margin}(dB).\end{aligned} \tag{45}$$

Once the cloud profile descriptors have been established, a statistical modeling of each one of its points (central box inside the population box of Fig. 6) is synthesized by fulfilling the cloud profile. This is done by selecting at random for each point i of the cloud, the following descriptors:

(1) The uncertain point centers, (that lie exactly on the law). This selection requires choosing previously a set φ_i of n angles uniformly inside the flare angle interval $[\varphi_{\min}, \varphi_{\max}]$. Once the φ_i have been chosen, the coordinates of data centres (x_i, y_i) can be evaluated by (46):

$$\begin{cases} x_i = D_{true}\cos(\Psi_{true}) - D_{true}\sin(\Psi_{true})\tan(\varphi_i)\\ y_i = D_{true}\sin(\Psi_{true}) + D_{true}\cos(\Psi_{true})\tan(\varphi_i).\end{cases} \tag{46}$$

(2) The value of the inter-axial cross correlations ρ_{xyi} associated with each uncertain point. These are chosen uniformly inside the cloud profile interval $[\rho_{\min}, \rho_{\max}]$.

(3) The value of the directional deviation of error on X and Y directions, σ_{xi}, σ_{yi} associated with each uncertain point. These values are choosing at random according with a log-uniform distribution inside the ranges $[\sigma x_{\min}, \sigma x_{\max}]$ and $[\sigma y_{\min}, \sigma y_{\max}]$. In fact, this is done by choosing $\sigma_{xi}(dB)$, $\sigma_{yi}(dB)$ uniformly inside $[\sigma x_{\min}(dB), \sigma x_{\max}(dB)]$ and $[\sigma y_{\min}(dB), \sigma y_{\max}(dB)]$, and then getting σ_{xi}, σ_{yi} by simply converting $\sigma_{xi}(dB)$, $\sigma_{yi}(dB)$ to linear units by

$$\begin{aligned}\sigma_{xi}(dB) &= 10^{\frac{\sigma_{xi}(dB)}{20}}\\ \sigma_{yi}(dB) &= 10^{\frac{\sigma_{yi}(dB)}{20}}.\end{aligned} \tag{47}$$

Now, one has a statistical description of each uncertain point of the cloud made up of five parameters: $\{x_i, y_i, \sigma_{xi}, \sigma_{yi}, \rho_{xyi}\}$. The last step (rightmost box inside the population box of Fig. 6) in the synthesis is to generate a noised sample $(\tilde{x}_i, \tilde{y}_i)$ for each point. For doing this, some noise $(\varepsilon x_i, \varepsilon y_i)$ is added (48) to the data centers (x_i, y_i).

$$\begin{aligned} \tilde{x}_i &= x_i + \varepsilon x_i \\ \tilde{y}_i &= y_i + \varepsilon y_i. \end{aligned} \tag{48}$$

The noise is chosen at random, according with a zero mean normal bivariate distribution whose covariance matrix for each point Σ_i is given by (49):

$$\Sigma_i = \begin{pmatrix} \sigma_{xi}^2 & \rho_{xyi}\sigma_{xi}\sigma_{yi} \\ \rho_{xyi}\sigma_{xi}\sigma_{yi} & \sigma_{yi}^2 \end{pmatrix}. \tag{49}$$

At this point, the set of descriptors $\{\tilde{x}_i, \tilde{y}_i, \sigma_{xi}, \sigma_{yi}, \rho_{xyi}\}$ of the cloud (Fig. 7), are fully defined in a data structure that can be supplied to a computer regression method in order to evaluate its regression estimated line $r_{guess} \equiv (D_{gues}, \Psi_{gues})$.

Next, these data (uncertain point realization) are supplied to three regression methods, namely the regression method under evaluation (PDIM), and another two classical methods that serve as contrast references. The chosen contrast methods are WLS (Asuero and González, 2007; Mandel and McCrackin, 1988) that considers null variances on X axis, and the reciprocal of WLS (that is the method obtained by interchanging the X and Y axis roles in WLS). Both reference methods are analytically solvable and their behavior and features are well known.

The results of the regression methods on the population of clouds are stored in a history file and a merit function of the regression quality is calculated and presented.

5.2. *Results*

Once the trail synthesis and evaluation procedures have been established, some simulations were performed in order to show PDIM performance with respect to the reference methods under some selected conditions.

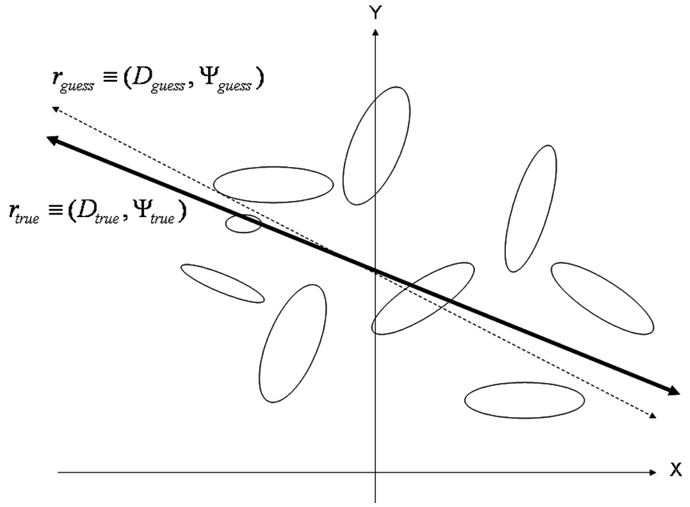
Y
$r_{guess} \equiv (D_{guess}, \Psi_{guess})$
$r_{true} \equiv (D_{true}, \Psi_{true})$
X

Figure 7. Uncertainty normal points, r_{true} and r_{guess}.

All tests in this section were obtained following steps below.

1. Imposing the fixed cloud descriptors values according with the selected conditions, and the number of trial clouds T we are going to use for the estimation at each point of the graph.
2. Looping the variable descriptor values, i.e. the values of the cloud descriptors whose impact on the performance is going to be evaluated.
3. Calculating the merit function of interest by using the history file $r_{guess}\{t\}$ and the previous known goal law r_{true}.
4. Displaying the merit function on the vertical axis as a function of the variable parameter on the horizontal axis.

5.2.1. *Robustness and Angular Independence.* These tests (Figs. 8 and 9) analyze the compared performance of the three methods against the angle descriptor: ψ_{true}. ψ_{true} is presented on the horizontal axis of the figures, in turns units. The choices for the other parameters are shown in Table 2.

Figure 8 shows on the vertical axis the mean quadratic dispersion of error that ψ_{guess} commit with respect to ψ_{true} in logarithm units. The displayed error measure is defined as:

$$\Psi \text{ dispersion (dB/Radian)} = 10\,Log_{10}\frac{\sum_{t=1}^{T}\left(\psi_{guess}\{t\} - \psi_{true}\right)^2}{T}. \qquad (50)$$

Figure 9 shows in ordinates the mean quadratic dispersion of the error that $D_{calculated}$ commit with respect to D_{true} straight line descriptor in logarithm units.

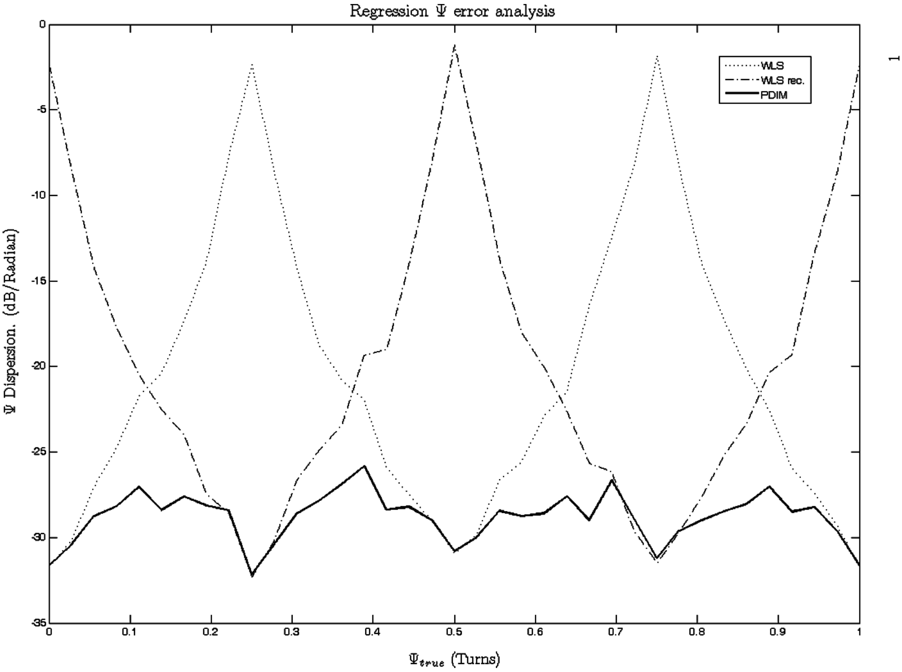

Figure 8. Angular independence. Ψ error analysis.

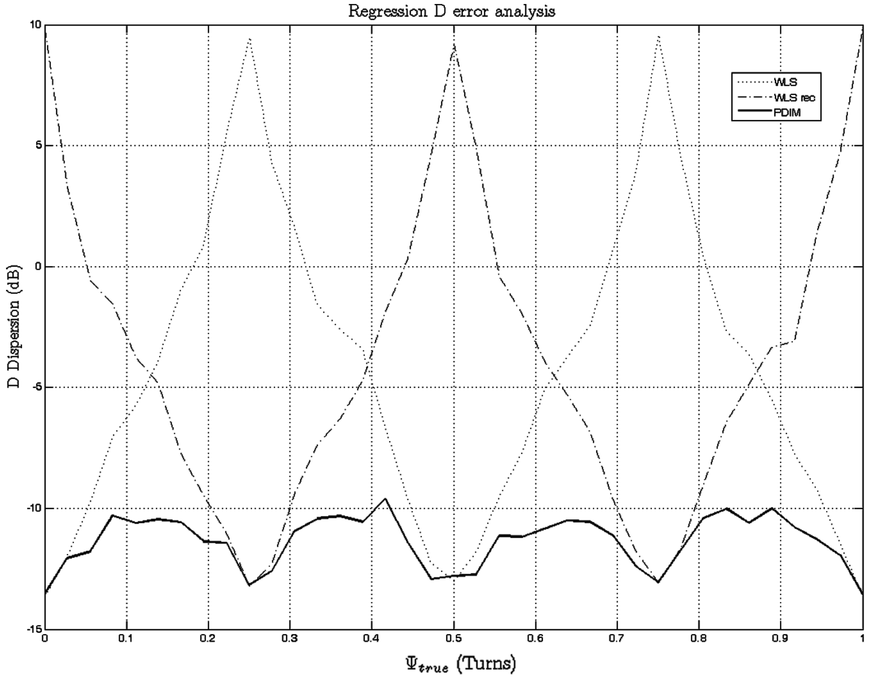

Figure 9. Angular independence. *D* error analysis.

The displayed error measure is defined as:

$$D \text{ dispersion (dB)} = 10\,Log_{10}\frac{\sum_{t=1}^{T}(D_{guess}\{t\} - D_{true})^2}{T}. \tag{51}$$

Both Figs. 8 and 9 show that PDIM is at least as good as the best of the two reference methods, for any value of the angle ψ_{true}. For horizontal (vertical) straight lines, X-axis deviations (Y-axis deviations) have no influence on PDIM results so in these cases the univariate WLS (reciprocal WLS) just reach PDIM

Table 2
Concrete function chosen for the tests of angular independence

Fixed parameter meaning	Fixed parameter symbol	Fixed parameter value
Cardinal of trials	T	300
Flare angle interval	$[\varphi_{\min}, \varphi_{\max}]$	$[-0.8\frac{\pi}{2}, 0.8\frac{\pi}{2}]$Radians
Distance true straight line descriptor	D_{true}	8
Cardinal of data's cloud	n	12
Correlation interval	$[\rho_{\min}, \rho_{\max}]$	[0,0]
Deviation minima	$\sigma x_{\min}(dB), \sigma y_{\min}(dB)$	0dB,0dB
Deviation margins	$\sigma x_{m\,\arg in}(dB), \sigma y_{\min}(dB)$	12dB,12dB

results. For other ψ_{true} values, the performance of the PDIM improves significantly any of the reference methods results.

Although PDIM is fully isotropic, a residual 3dB fluctuation of performance remains in both figures. This does not come from PDIM but from some bias that has its origin in the trial cloud synthesis methodology, due to the fact that deviation minima and margins are referred to a specific axes descriptive system (X, Y).

5.2.2. *Performance Analysis Against Data Deviation Disparity.* Now, we will analyze the compared performance of the PDIM with respect to the reference univariate methods (WLS and reciprocal WLS) as the disparity between the deviations of the data (the heteroscedasticity degree) grows.

The horizontal axis of Fig. 10 represents the heteroscedasticity degree through the value $\sigma_{\text{margin}}(dB)$ that we assign to the directional deviation freedom margins along both X and Y data axes. That is, we set:

$$\sigma x_{m\arg in}(dB) = \sigma_{m\arg in}(dB)$$
$$\sigma y_{m\arg in}(dB) = \sigma_{m\arg in}(dB)$$

for each value of $\sigma_{m\arg in}(dB)$.

On the vertical axis, Fig. 10 shows the mean quadratic dispersion of the error that ψ_{guess} commit in logarithm units (52):

$$\Psi \text{ dispersion (dB/Radian)} = 10\,Log_{10}\frac{\sum_{t=1}^{T}(\psi_{guess}\{t\} - \psi_{true})^2}{T}. \qquad (52)$$

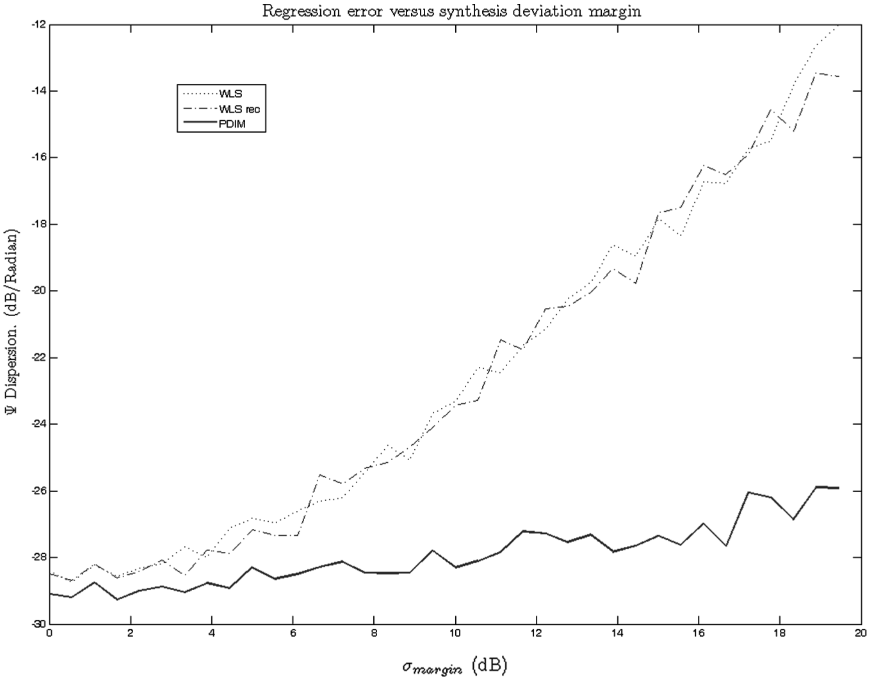

Figure 10. Performance against data deviation disparity.

Table 3
Concrete function chosen for the tests of performance against data deviation disparity

Fixed parameter meaning	Fixed parameter symbol	Fixed parameter value
Cardinal of trials	T	1000
Flare angle interval	$[\varphi_{\min}, \varphi_{\max}]$	$[-0.8\frac{\pi}{2}, 0.8\frac{\pi}{2}]$Radians
Distance true straight line descriptor	D_{true}	8
Angle true straight line descriptor	Ψ_{true}	$\pi/4$
Cardinal of cloud's data	n	12
Correlation interval	$[\rho_{\min}, \rho_{\max}]$	[0,0]
Deviation minima	$\sigma x_{\min}(dB), \sigma y_{\min}(dB)$	0dB,0dB

The choices for the others trial parameters are shown in Table 3.

5.2.3. *Performance Analysis Against Interaxial Cross-Correlation Data Disparity.* Finally, in Fig. 11, we analyze the ability of PDIM using the knowledge of the interaxial correlation coefficient ρ_{xyi}. As univariate methods are blind to this information, it is expected to have a performance improvement.

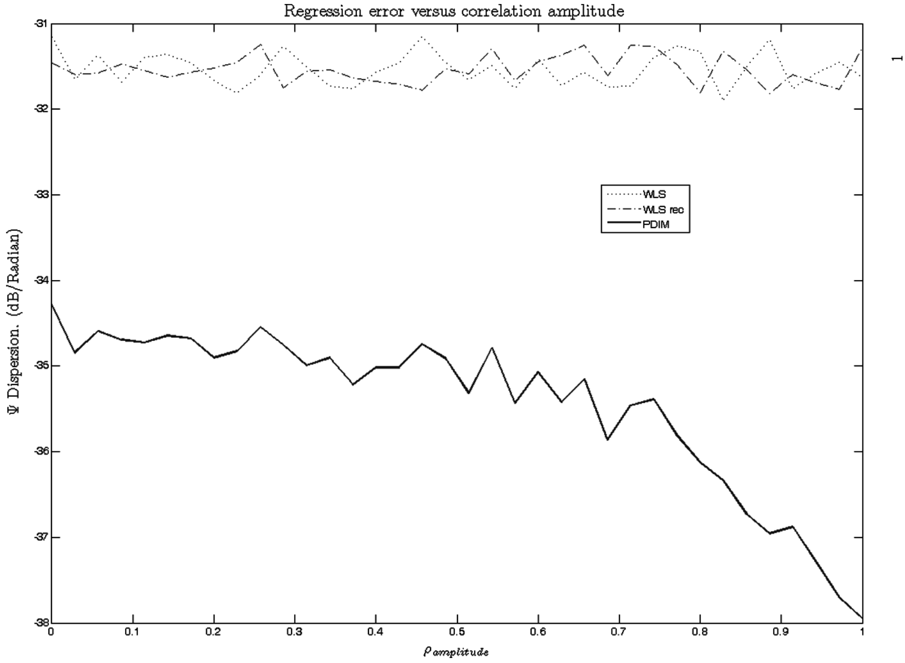

Figure 11. Compared performance in presence of inter-axial dependence.

Table 4
Concrete function chosen for the tests of performance against interaxial disparity

Fixed parameter meaning	Fixed parameter symbol	Fixed parameter value
Cardinal of trials	T	1000
Flare angle interval	$[\varphi_{\min}, \varphi_{\max}]$	$[-0.8\frac{\pi}{2}, 0.8\frac{\pi}{2}]$Radians
Distance straight line descriptor	D_{true}	8
Angle straight line descriptor	Ψ_{true}	$\pi/4$
Cardinal of data's cloud	n	12
Deviation minima	$\sigma x_{\min}(dB), \sigma y_{\min}(dB)$	6dB,6dB
Deviation margins	$\sigma x_{m\arg in}(dB), \sigma y_{\min}(dB)$	0dB,0dB

In the test, the interval of allowed values of ρ_{xyi}, $[\rho_{\min}, \rho_{\max}]$ was selected by the formula

$$[\rho_{\min}, \rho_{\max}] = [-\rho_{amplitude}, \rho_{amplitude}],$$

Here, $\rho_{amplitude}$ is the inter-axial correlation amplitude freedom that is the variable that has been represented on the horizontal axis of Fig. 11. On the vertical axis, the same performance measure (52) of former results has been represented.

As inter-axial correlation amplitude freedom grows, it is expected that the performance also grows, as effectively occurs in the graph. It can also be observed that even for the $\rho_{amplitude} = 0$ value, there is a 3dB gain of PDIM against univariate methods. This 3dB improvement is due to the advantage that gives to PDIM the knowledge of the deviations on the two axes.

6. Conclusions

When data are scarce or experiments are expensive, it is important to take advantage of all the available knowledge about the quality of the measurements by using sophisticated regression methods in order to improve the accuracy of the results.

In this article, we presented a regression methodology that is able to convert any statistically formulated regression problem into a scalar optimization. It can deal with arbitrary data error distribution functions and can be particularized, by making some simple additional assumptions in order to mimic classic regression methods as TLS, WLS, or BLS by a single algorithm.

The methodology is based on the definition of a measure (TRDND) on the data space and the use of polar descriptors (D, ψ) for the straight lines, instead of the usual "slope, y-intercept" (m, b) descriptors. Polar descriptors avoid overflows for vertical lines, and isolate regression results from the concrete selection of axes on which the data are given.

Finally, a practical algorithm: PDIM was also developed in order to solve the problem that arises when the methodology is applied to normal clouds. PDIM is able to advantageously afford the most complex cases of bivariate heteroscedastic data with inter-axial dependence.

PDIM was tested and compared against others heteroscedastic univariate methods that have singularities when laws are vertical or horizontal straight lines. It avoids these singularities and offers the best performance irrespective of the regression line slope.

PDIM results have been found to be especially profitable when dealing with data that have a strong heteroscedasticity degree affecting both components and/or with data subject to inter-axial dependence.

References

Asuero, G., González, G. (2007). Fitting straight lines with replicated observations by linear regresión. III. Weighting data. *Crit. Rev. Analyt. Chem.* 37:143–172.

Cheng, C., Riu, J. (2006). on estimating linear relationships when both variables are subject to heteroscedastic measurement errors. *Technometrics* 48:511–519.

Duda, R. O., Hart, P. E., Store, D. G. (1997). *Pattern Classification*. New York: Wiley Interscience.

González, J. R., Vázquez, M., Núñez, N., Algora, C., Rey-Stolle, I., Galiana, B. (2009). Reliability analysis of temperature step-stress tests on III–V high concentrator. *Microelectronics Reliability* 49:673–822.

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proc. Nat. Instit. Sci. India* 12:49–55.

Mandel, J., McCrackin, F. L. (1988). An iterative self-weighting procedure for fitting straight lines to heteroscedastic data. *Commun. Statist. Simul. Computat.* 17(2):609—635.

Markovsky, I., Van Huffel, S. (2007). Overview of total least-squares methods. *Signal Process.* 87:2283–2302.

Martínez, A., del Río, F. J., Riu, J., Rius, F. X. (1999). Detecting proporcional and constant bias in method comparison studios by using linear regression with errors in both axes. *Chemometr. Intelligent Lab. Syst.* 49:181–195.

Rawlings, J. O., Pantula, S. G., Dickey, D. A. (1998). *Applied Regression Analysis: A Research Tool*. 2nd ed. New York: Springer Text in Statistics.

Sayago, A., Boccio, M., Asuero, A. G. (2004). fitting straight lines with replicated observations by linear regresión: the least squares postulates. *Crit. Rev. Analyt. Chem.* 34:39–50.

Van Huffel, S., Cheng, C., Mastronardi, N., Paige, C., Kukush, A. (2007). Total least squares and errors-in-variables modelling. *Computat. Statist. Data Anal.* 52:1076–1079.

Vázquez, M., Algora, C., Rey-Stolle, I., Algora, C. (2007). III–V concentration solar cell reliability prediction based on quantitative led reliability data. *Prog. Phot. Res. Appl.* 15:477–491.

Wilcox, R. R. (2009). Robust multivariate regression when there is heteroscedasticity. *Commun. Statist. Simul. Computat.* 38(1):1–13.

Yu, Q., Chappell, R., Wong, G. Y. C., Hsu, Y., Mazur, M. (2008). Relationship between the cox, lehmann, weibull, and accelerated lifetime models. *Commun. Statist. Theor. Meth.* 37(9):1458–1470.