Multiple Linear Regression for Histogram Data using Least Squares of Quantile Functions: a Two-components model.

Rosanna Verde*, Antonio Irpino*

* Dipartimento di Scienze Politiche "J. Monnet"
Seconda Universitá degli Studi di Napoli
Viale Eliitico 31, Caserta, Italy
rosanna.verde@unina2.it, antonio.irpino@unina2.it

Abstract. Histograms are commonly used for representing summaries of observed data and they can be considered non parametric estimates of probability distributions. Symbolic Data Analysis formalized the concept of *histogram symbolic variable*, as a variable which allows to describe statistical units by histograms instead of single values. In this paper we present a linear regression model for multivariate histogram variables. We use a Least Square estimation method where the sum of squared errors is defined according to the ℓ_2 Wasserstein metric between the observed and the predicted histogram data. Consistently with the l_2 Wasserstein metric, we solve the Least Square computational problem by introducing a suitable inner product between two vectors of histogram data. Finally, measures of goodness of fit are discussed and an application on real data shows some interpretative advantages of the proposed method.

1 Introduction

Symbolic Data Analysis (SDA) is a relatively new statistical approach concerning the analysis of *higher level individuals* (like typologies, classes or concepts) that are described by *multi-valued variables* (Bock and Diday (2000); Diday and Noirhomme-Fraiture (2008)).
The term *symbolic variable* was coined in order to introduce such new set-valued descriptions. In a classic data table ($n \times p$ individuals per variables) each individual is described by a vector of values, similarly, in a *symbolic data table* each individual is described by a vector of set-valued descriptions (like intervals of values, histograms, set of numbers or of categories, sometimes equipped with weights, probabilities, frequencies, an so on). According to the taxonomy of symbolic variables presented in Bock and Diday (2000) we may consider as numerical symbolic variables all those symbolic variables whose support is numeric. The main quantitative and multi-valued symbolic variable types are *interval* and *modal numeric symbolic variable*.
Linear regression models allow to modeling the linear relationship between a quantitative response *symbolic* variable and a set of independent or explicative quantitative *symbolic* variables of the same type. Regression models for interval data extend the classic linear model to interval variables (see Afonso et al. (2008) and Lima Neto and de Carvalho (2010)) and the therein

references for a full overview of regression models for interval *symbolic* data).
In this paper we focus on the regression analysis of *histogram symbolic data* which are realizations of *histogram symbolic variables* (such variables are particular cases of modal symbolic variables). In Billard and Diday (2006) was presented a first regression model for histogram variables. This approach is based on the basic statistics (i.e., the mean, the standard deviation and the correlation) proposed by Bertrand and Goupil (2000). However, the proposed model allows to predict punctual values and only using Monte Carlo procedures it is possible to predict a response histogram variable given a set of observed explicative histogram variables (the authors themselves leave as open problem the output prediction in terms of symbolic data). In order to be fulfill such requirement, (i.e. a model that have in input histogram variables and as output a response histogram variable too) Verde and Irpino (2010) proposed a simple linear regression model for histogram variables. The parameters of the model are estimated using the ℓ_2 Wasserstein distance (also known as Mallow's distance (Rüschendorf, 2001)) for defining the residual sum of square criterion of the Least Square method. Recently, Dias and Brito (2011) proposed a multivariate linear regression model for histogram data (HD) based on the ℓ_2 Wasserstein distance and on a constrained Least Square optimization problem. The authors proposed a regression model with a doubled number of predictors including, for each independent variable, its *symmetric* histogram variables.
In the present paper, we propose an extension of the simple model of Verde and Irpino (2010) to the multivariate case. We do not introduce new variables (like in Dias and Brito (2011)) but we use a particular decomposition of the observed variables that both, fits the data and allows an easy interpretation of the estimated parameters of the model. Observing that ℓ_2 Wasserstein distance is computed using the quantile functions (*qf*s) associated with histograms, we show how to compute the inner product between two vectors of *qf*s according to a decomposition of the Wasserstein metric introduced by Irpino and Romano (2007). We also furnish a novel goodness of fit index for the evaluation of the model.
The paper is organized as follows: the section 2 introduces the histogram variables according to the symbolic variables definition given in Bock and Diday (2000) and Diday and Noirhomme-Fraiture (2008). Considering the ℓ_2 Wasserstein metric, we present the derived main statistics and algebraic operators for histogram data. In section 3 the regression model for histogram symbolic variables is introduced and detailed as well as the main goodness of fit indices for the evaluation of the model. Finally, section 4 compares our proposed model to the other ones presented in the literature using a climatic dataset and highlights some interpretative results.

2 Histogram variables: definitions and basic statistics

Let us $E = \{e_1, \ldots, e_n\}$ be a set of n statistical units (individuals, concepts, classes). According to Bock and Diday (2000) a *symbolic modal variable* X, with domain D, is a mapping $X : E \to P = M(D) \in \Re^+$ from E into the family of all non-negative measures (frequency, probability or weight distributions) on the domain D. For each element e_i of E a modal variable assigns a measure $M(I_i) = P_i$ defined on the set of values $I_i \subseteq D$.

If $X(e_i)$ has the same properties of a random variable, it is defined as a *Numerical Probabilistic* (Modal) *Symbolic Variable* and P_i can be a probability density function, a histogram, an empirical frequency distribution $f_i(x)$.

In this paper we refer only to *Numerical Probabilistic (Modal) Symbolic Data*, that are in the domain of *Numerical Probabilistic* (Modal) *Symbolic Variables*, and in particular we refer to *Histogram Data* (HD), that are in the domain of *Histogram Variables* (HV). In order to simplify the notation, hereafter we denote $X(e_i)$ with $X(i)$ (for $i = 1, \ldots, n$) , i.e. the modal description of the individual e_i for the symbolic variable X. In this paper, we assume that $X(i)$ (being X a HV) is a histogram partitioned into h_i bins I_{ui} ($u = 1, \ldots, h_i$) and it is described as the following vector of pairs:

$$X(i) = \{(I_{1i}, p_{1i}), ..., (I_{ui}, p_{ui}), ..., (I_{h_i i}, p_{h_i i})\},$$

where $I_{ui} = [a_{ui}, b_{ui}[$ is an interval, given $u \neq u'$ then $I_{ui} \cap I_{u'i} = \emptyset$ and $0 \leq p_{ui} \leq 1$ such that $\sum_{u=1}^{h_i} p_{ui} = 1$, and $u = 1, \ldots, h_i$.

Basic statistics for HD In general, a histogram (Pearson, 1895) can be considered as a simplified non-parametric estimates of a probability distribution. As usual when dealing with histograms, each bin is uniformly distributed and then for each $X(i)$ we can define a *pdf* density function, a *cdf* distribution function and the corresponding quantile function (*qf*). The *pdf* of $X(i)$ is:

$$f_i(x) = \begin{cases} \dfrac{p_{ui}}{b_{ui} - a_{ui}} & if \quad x \in I_{ui} \\ 0 & otherwise \end{cases} \tag{1}$$

We here introduce the quantities w_{ui} that represent the *cdf* values observed at the upper extreme of each bin of $X(i)$ as follows:

$$w_{0i} = 0; \quad w_{ui} = \sum_{\ell=1}^{u} p_{\ell i} \qquad u = 1, \ldots, h_i. \tag{2}$$

in this way it is easy to describe the *cdf* associated with $X(i)$ as:

$$F_i(x) = \begin{cases} 0 & if \quad x < a_{1i} \\ w_{(u-1)i} + p_{ui} \dfrac{x - a_{ui}}{b_{ui} - a_{ui}} & if\ x \in I_{ui}\ and\ u > 0 \\ 1 & if \quad x > b_{h_i i} \end{cases} \tag{3}$$

The *cdf* of a histogram is a piece-wise linear function, consequently also the corresponding *qf* is piece-wise linear. Given $t \in [0, 1]$, we denote with $q_i(t)$ the $t - th$ quantile of $X(i)$ and is expressed as follows:

$$q_i(t) = F_i^{-1}(t) = a_{ui} + \frac{t - w_{(u-1)i}}{w_{ui} - w_{(u-1)i}} (b_{ui} - a_{ui}) \quad 0 \leq w_{(u-1)i} \leq t \leq w_{ui} \leq 1. \tag{4}$$

Using the proposed notation we recall how to compute the basic statistics for each histogram. We denote with $\bar{x}_i$ the mean and with s_i the standard deviation associated with $X(i)$. Let $c_{\ell i} = (a_{\ell i} + b_{\ell i})/2$ and $r_{\ell i} = (b_{\ell i} - a_{\ell i})/2$ be the center and the radius of the ℓ-th bin of $X(i)$, we recall that the mean and the standard deviation of a histogram can be calculated, in a linear

time w.r.t. the number of bins h_i of the histogram $X(i)$, as follows:

$$\bar{x}_i = \int_{a_1}^{b_{h_i}} x f_i(x) dx = \int_{a_1}^{b_{h_i}} x dF_i(x) = \int_0^1 q_i(t) dt = \sum_{\ell=1}^{h_i} p_{\ell i} c_{\ell i}. \tag{5}$$

while the standard deviation is:

$$s_i = \left[\int_{a_1}^{b_{h_i}} [x - \bar{x}_i]^2 f_i(x) dx \right]^{\frac{1}{2}} = \left[\int_0^1 [q_i(t) - \bar{x}_i]^2 dt \right]^{\frac{1}{2}} = \left[\sum_{\ell=1}^{h_i} p_{\ell i} \left(c_{\ell i}^2 + \frac{1}{3} r_{\ell i}^2 \right) - \bar{x}_i^2 \right]^{\frac{1}{2}}. \tag{6}$$

Basic statistics for histogram variables according to the ℓ_2 Wasserstein metric. Verde and Irpino (2007) compared different probabilistic metrics for histogram data. The ℓ_2 Wasserstein distance (known also as Mallow's distance) showed some useful properties for the definition of basic statistics of histogram variables. According to the notation presented in paper, the ℓ_2 Wasserstein distance between two histograms $X(i)$ and $X(j)$ is:

$$d_W\left(X(i), X(j)\right) = \sqrt{\int_0^1 \left[q_i(t) - q_j(t)\right]^2 dt} \tag{7}$$

The equation (7) implies the invertibility of the *cdf*s for expressing the distance into a closed form and this is not in general true. Fortunately, for histograms we can define a closed form of the distance. According to Irpino et al. (2006), to compute the distance between two histograms $X(i)$ and $X(j)$, exactly and in a finite number of steps, we need to identify a set of uniformly dense intervals to be compared on the basis of the two *qf*s. In order to find such set of intervals, firstly we merge the cumulated weights of the two histograms into a single vector:

$$\mathbf{v} = \left[w_{0i}, \ldots, w_{ui}, \ldots, w_{h_i i}, w_{0j}, \ldots, w_{vi}, \ldots, w_{h_j j} \right] \tag{8}$$

where h_i and h_j are respectively the number of bins of $X(i)$ and $X(j)$. After that, we construct the vector $\mathbf{w}$ that contains the sorted and unique values of w:

$$\mathbf{w} = [w_0, ..., w_\ell, ..., w_m] \tag{9}$$

where $w_0 = 0$, $w_m = 1$ and $\max(h_i, h_j) \leq m \leq (h_i + h_j - 1)$.
With the same vector, it is possible to associate a vector of m weights $\pi = [\pi_\ell]$ where $\pi_\ell = w_\ell - w_{\ell-1}$, and the d_W^2 between two histograms can be expressed as follows::

$$d_W^2(X(i), X(j)) = \sum_{\ell=1}^{m} \int_{w_{\ell-1}}^{w_\ell} \left(q_i(t) - q_j(t)\right)^2 dt. \tag{10}$$

Each pair $(w_{\ell-1}, w_\ell)$ allows us to identify two uniform intervals, one for $X(i)$ and one for $X(j)$, having respectively the following bounds that are computed using their quantile functions:

$$I_{\ell i}^* = [q_i(w_{\ell-1}); q_i(w_\ell)] \quad \text{and} \quad I_{\ell j}^* = [q_j(w_{\ell-1}); q_j(w_\ell)]. \tag{11}$$

The asterics indicate that we are not considering the original bins but a new set of bins which arise by partitions of the original ones: i.e. we describe the same histogram using more bins but without modifying the density function. For each bin, it is possible to compute the centers and the radii as follows:

$$\begin{array}{ll} c^*_{\ell i} = (q_i(w_\ell) + q_i(w_{\ell-1}))/2 & r^*_{\ell i} = (q_i(w_\ell) - q_i(w_{\ell-1}))/2 \\ c^*_{\ell j} = (q_j(w_\ell) + q_j(w_{\ell-1}))/2 & r^*_{\ell j} = (q_j(w_\ell) - q_j(w_{\ell-1}))/2. \end{array}$$

Finally, considering the m bins as m uniform distributions the squared ℓ_2 Wasserstein distance between two HD's is:

$$d^2_W(X(i), X(j)) := \sum_{\ell=1}^{m} \pi_\ell \left[\left(c^*_{\ell i} - c^*_{\ell j}\right)^2 + \frac{1}{3}\left(r^*_{\ell i} - r^*_{\ell j}\right)^2 \right]. \tag{12}$$

Equation (12) (Irpino et al., 2006) allows to define an inertia measure of a set of HD and the *mean histogram* as this histogram which minimizes such inertia. Having observed n units described by the histogram variable X, the *mean histogram* is the histogram associated to the quantile function $\bar{q}(t)$, which minimizes the following sum of the squared differences:

$$SS(X) = \sum_{i=1}^{n} d^2_W(X(i), \bar{X}) = \sum_{i=1}^{n} \int_0^1 \left[q_i(t) - \bar{q}(t)\right]^2 dt \tag{13}$$

The minimum of $SS(X)$ is achieved when $\bar{q}(t) = \sum_{i=1}^{n} q_i(t)/n$ for each $t \in [0,1]$. Also in this case it possible to define the quantile function $\bar{q}(t)$ in a linear (w.r.t. the sum of the number of bins of the HD's) time. By the vector

$$\mathbf{v} = [w_{01}, \ldots, w_{h_1 1}, \ldots, w_{0i}, \ldots, w_{h_i i}, \ldots, w_{0n}, \ldots, w_{h_n n}] \tag{14}$$

containing $n + \sum_{i=1}^{n} h_i$ elements and sorting the unique values, we obtain the vector $\mathbf{w}$

$$\mathbf{w} = [w_0, ..., w_\ell, ..., w_m] \tag{15}$$

where $min(h_i) \leq m \leq \left(\sum_{i=1}^{n} h_i - 2n + 1\right)$ is the number of bins of the *mean histogram* $\bar{X}$ associated with the $\bar{q}(t)$ *mean quantile function*:

$$\bar{X} = \left\{(\bar{I}_l, p_l) | l = 1, \ldots, m\right\} \tag{16}$$

where $\bar{I}_l = \left[\sum_{i=1}^{n} q_i(w_{(l-1)})/n; \sum_{i=1}^{n} q_i(w_l)/n\right]$ and $p_l = w_l - w_{l-1}$.
Equation (13) allows the definition of a standard deviation measure S_x for a histogram variable (Verde and Irpino, 2008) as follows:

$$S_x = \sqrt{SS(X)/n} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} d^2_W(X(i), \bar{X})}. \tag{17}$$

In order to simplify the notation of the presented formulas and for coherence with the notation generally used for denoting regression model variables,

- we denote with $x_{ij}(t)$ the quantile function $q_i(t)$ associated with the histogram description of the $i-th$ unit for the X_j independent histogram variable;
- we denote with $y_i(t)$ the quantile function associated to the histogram description of the $i-th$ unit for the Y dependent histogram variable.

Inner product between two histogram variables related to the ℓ_2 Wasserstein metric. Considering eq. (12) we express the inner product of two quantile functions associated with two HD as follows:

$$\langle x_i(t), x_j(t)\rangle = \int_0^1 x_i(t)\cdot x_j(t)dt = \sum_{\ell=1}^{m}\pi_\ell\left[c^*_{\ell i}\cdot c^*_{\ell j} + \frac{1}{3}r^*_{\ell i}\cdot r^*_{\ell j}\right]. \tag{18}$$

Given two HD's $X(i)$ and $X(j)$ and $x^c_f(t)$ and $x^c_g(t)$ the respective centered quantile functions, Cuesta-Albertos et al. (1997) showed that the ℓ_2 Wasserstein distance can be rewritten as

$$d^2_W(X(i),X(j)) = (\bar{x}_i-\bar{x}_j)^2 + \int_0^1\left[x^c_i(t)-x^c_j(t)\right]^2 dt. \tag{19}$$

This property allows us to consider the squared distance as the sum of two components: the first related to the locations of HD and the second related to their variability structure. An interesting decomposition of the ℓ_2 Wasserstein distance proposed by Irpino and Romano (2007) for continuous random variables is the following:

$$d^2_W(X(i),X(j)) = (\bar{x}_i-\bar{x}_j)^2 + (s_i-s_j)^2 + 2s_is_j\left(1-\rho(x_i,x_j)\right) \tag{20}$$

where $\rho(x_i,x_j)$ is the correlation coefficient between the two quantile functions, which, in the case of HD is:

$$\rho(x_i,x_j) = \frac{\int_0^1 x_i(t)x_j(t)dt - \bar{x}_i\cdot\bar{x}_j}{s_i\cdot s_j} = \frac{\sum_{\ell=1}^{m}\pi_\ell\left[c^*_{\ell i}c^*_{\ell j}+\frac{1}{3}r^*_{\ell i}r^*_{\ell j}\right]-\bar{x}_i\cdot\bar{x}_j}{s_i\cdot s_j}. \tag{21}$$

The equation (21) permits to write a general expression for the inner product between two *qf*s, where eq. (18) is a specific formula for HD, as follows:

$$\langle x_i(t), x_j(t)\rangle = \rho(x_i,x_j)\cdot s_i\cdot s_j + \bar{x}_i\cdot\bar{x}_j \tag{22}$$

that allows a better interpretation in term of scale, size and shape of the two HD's. Finally, given two vectors of quantile functions $\mathbf{x} = [x_i(t)]_{n\times 1}$and $\mathbf{y} = [y_i(t)]_{n\times 1}$, we can express the scalar product of two vectors of HD's as:

$$\mathbf{x^T y} = \int\sum_{i=1}^{n}(x_i(t),y_i(t))dt = \sum_{i=1}^{n}\left[\rho(x_i,y_i)\cdot s_{x_i}\cdot s_{y_i} + \bar{x}_i\cdot\bar{y}_i\right]. \tag{23}$$

It is worth noting that, when $\mathbf{x}$ is a vector of scalar, each value can be treated as a *impulse function*, i.e., $\bar{x}_i = x_i$ and $s_{x_i} = 0$, thus:

$$\mathbf{x^T y} = \sum_{i=1}^{n}\left[\rho(x_i,y_i)\cdot 0\cdot s_{y_i} + x_i\cdot\bar{y}_i\right] = \sum_{i=1}^{n}\left[x_i\cdot\bar{y}_i\right]. \tag{24}$$

3 Two components linear model: Least Squares estimation method for ℓ_2 Wasserstein based linear regression

Let us $X_1, \ldots, X_p$ be p independent histogram variables and Y a response histogram variable too. They are observed on a set E of n units. We represent them in a *symbolic data table* (where instead of a matrix of scalar values, it is a matrix of histogram data), that is:

$$[\mathbf{Y} \;\; \mathbf{X}] = [Y(i) \;\; X_1(i) \;\; \ldots \;\; X_p(i)]_{n\times 1}\,. \tag{25}$$

As in the classic regression approach we assume that the data are generated from the model:

$$\mathbf{Y} = \phi(\mathbf{X}|\beta) + \mathbf{e} \tag{26}$$

where e is a random error and $\phi(\mathbf{X})$ is linear with respect to the parameters β [1].
When the descriptors are HV, two main approaches for the estimation of the parameters of linear regression model have been proposed. Starting from the elementary statistics proposed by Bertrand and Goupil (2000), a first approach Billard and Diday (2006) extended of the classic OLS (Ordinary Least Squares) linear regression model to the histogram-valued variables. A second group of approaches is based on the use of the quantile functions (which are in bijection with their corresponding *pdf*'s) of the HD and of the ℓ_2 Wasserstein distance for defining the sum of square errors in the LS function. The idea behind the latest approaches is to predict a quantile function after having observed a set of quantile functions as predictors. In this paper, we follow the latest approach.

Given the matrix (25), we consider the associated matrix $\mathbf{M}$ containing the corresponding quantile functions:

$$\mathbf{M} = [\mathbf{Y} \;\; \mathbf{X}] = [y_i(t) \;\; x_{i1}(t) \;\; \ldots \;\; x_{ip}(t)]_{n\times(p+1)} \tag{27}$$

A natural choice for the linear model should be:

$$y_i(t) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}(t) + e_i(t) \quad \forall t \in [0;1],$$

and the Sum of Squared Errors (SSE) criterion to minimize for the solution of the OLS problem is related to the Squared ℓ_2 Wasserstein distance is:

$$SSE = \sum_{i=1}^{n} \int_0^1 [e_i(t)]^2\, dt = \sum_{i=1}^{n} d_W^2 \left(y_i(t), \left[\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}(t) \right] \right) =$$

$$= \sum_{i=1}^{n} \int_0^1 \left[y_i(t) - \left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}(t) \right) \right]^2 dt.$$

A problem arises for the linear combination of quantile functions. Quantile functions are not decreasing function in $[0;1]$, thus only if $\beta_j \geq 0$ $(j = 1, \ldots, p)$ it assures that $y_i(t)$ is a

1. We assume the random components of the model $e_i(t)$ as generated by an error function.

quantile function too. To tackle this issue, Verde and Irpino (2010) presented a new regression model based on the decomposition of d_W^2 for the simple regression model, while Dias and Brito (2011) proposed to consider an adjoint set of histogram variables related to the observed ones and a constrained LS method.
In this paper, we consider a multivariate extension of the model proposed by Verde and Irpino (2010). Considering that $x_{ij}^c(t) = x_{ij}(t) - \bar{x}_{ij}$, each element of $\mathbf{X}$, as well as each element of $\mathbf{Y}$, can be rewritten as:

$$x_{ij}(t) = x_{ij}^c(t) + \bar{x}_{ij}.$$

The matrix $\mathbf{M}$ is then transformed as:

$$\mathbf{M} = \left[\bar{\mathbf{Y}} + \mathbf{Y^c} \;\; \bar{\mathbf{X}} + \mathbf{X^c}\right] = \left[\bar{\mathbf{Y}} \;\; \bar{\mathbf{X}}\right] + \left[\mathbf{Y^c} \;\; \mathbf{X^c}\right] \tag{28}$$

where:
- $\bar{\mathbf{Y}} = [\bar{y}_i]_{n\times 1}$ is the vector of the means of the $f_i(y)$,
- $\mathbf{Y}^c = [y_i^c(t)]_{n\times 1}$ is the vector of the centred quantile functions of $f_i(y)$'s,
- $\bar{\mathbf{X}} = [\bar{x}_{ij}]_{n\times p}$ is the matrix of the means of the $f_i(x_j)$,
- $\mathbf{X^c} = \left[x_{ij}^c(t)\right]_{n\times p}$ is the matrix of the centred quantile functions of $f_i(x_j)$'s.

We consider each quantile function $y_i(t)$ as a linear combination of the means $\bar{x}_{ij}$ (that are scalars) and of the centered quantile functions $x_{ij}^c(t)$ plus a function error term $e_i(t)$ as:

$$y_i(t) = \beta_0 + \sum_{j=1}^{p} \beta_j \bar{x}_{ij} + \sum_{j=1}^{p} \gamma_j x_{ij}^c(t) + e_i(t) \tag{29}$$

Denoting $\bar{\mathbf{X}}_+ = [\mathbf{1}|\bar{\mathbf{X}}]$, we rewrite the model in (29) using the following matrix notation:

$$\mathbf{Y} = \bar{\mathbf{X}}_+\mathbf{B} + \mathbf{X}\Gamma + \mathbf{e}. \tag{30}$$

The model in eq. (30) permits to interpret the relationship between the predictors and the dependent variable taking into consideration two main aspects: the $\mathbf{B}$ vector quantifies the linear relationship due to the positions of the HD (their means), while the Γ vector quantifies the effect of the internal variability of the predictor variables X_j on the response variable Y. Using the ℓ_2 Wasserstein distance in the LS estimation method, we write the SSE function in scalar form as follows:

$$SSE(\mathbf{B}, \Gamma) = \sum_{i=1}^{n} \int_0^1 \left[y_i(t) - \beta_0 - \sum_{j=1}^{p} \beta_j \bar{x}_{ij} - \sum_{j=1}^{p} \gamma_j x_{ij}^c(t) \right]^2 dt \tag{31}$$

while in matrix form is:

$$SSE(\mathbf{B}, \Gamma) = \mathbf{e^T e} = \left[\mathbf{Y} - \bar{\mathbf{X}}_+ B - \mathbf{X}^c\Gamma\right]^T \left[\mathbf{Y} - \bar{\mathbf{X}}_+ B - \mathbf{X}^c\Gamma\right]. \tag{32}$$

Considering the algebraic operators introduced in eq. (23) and (22) follows that:

$$\bar{\mathbf{X}}_+^T\mathbf{X^c} = \mathbf{0}_{(p+1)\times p}, \;\; \bar{\mathbf{X}}_+^T\mathbf{Y} = \bar{\mathbf{X}}_+^T\bar{\mathbf{Y}}, \;\; \mathbf{X^{c}}^T\mathbf{Y} = \mathbf{X^{c}}^T\mathbf{Y^c} \tag{33}$$

Thus, $SSE(\mathbf{B}, \Gamma)$ can be decomposed into two positive quantities as follows:

$$SSE(\mathbf{B}, \Gamma) = SSE(\mathbf{B}) + SSE(\Gamma) = \bar{\mathbf{e}}^T\bar{\mathbf{e}} + \left(\mathbf{e^c}\right)^{\mathbf{T}}\mathbf{e^c} \tag{34}$$

being:

$$\bar{\mathbf{e}} = \bar{\mathbf{Y}} - \bar{\mathbf{X}}_{+}\mathbf{B} \qquad \mathbf{e^c} = \mathbf{Y}^c - \mathbf{X}^c\Gamma \tag{35}$$

with $\bar{\mathbf{e}} = [\bar{\mathbf{e}}_i]_{n\times 1}$ a vector of real values. We may express the single minimization problem as the minimization of two independent functions: the first one related to the means of the predictor quantile functions $\bar{x}_{ij}$'s in $\bar{\mathbf{X}}_{+}$, and the second one related to the variability of the centered quantile distributions $x^c_{ij}(t)$'s in $\mathbf{X^c}$. Then two models are independently estimated:

$$\bar{\mathbf{Y}} = \bar{\mathbf{X}}_{+}\mathbf{B} + \bar{\mathbf{e}} \qquad \mathbf{Y^c} = \mathbf{X}^c\Gamma + \mathbf{e}^c \tag{36}$$

The first equation is solved as classical OLS problem for the estimation of $\mathbf{B}$:

$$\hat{\mathbf{B}} = \left(\bar{\mathbf{X}}_{+}^T\bar{\mathbf{X}}\right)^{-1}\bar{\mathbf{X}}_{+}^T\bar{\mathbf{Y}}. \tag{37}$$

The second equation in (36) is solved using the NNLS (Non Negative Least Squares) algorithm proposed by Lawson and Hanson (1974) and the inner product as defined in eq. (18) and (23) for the matrix operations:

$$\begin{array}{rcl} argmin_{\Gamma}\, SSE(\Gamma) & = & \left[\mathbf{Y}^c - \mathbf{X}^c\Gamma\right]^T\left[\mathbf{Y}^c - \mathbf{X}^c\Gamma\right] \\ s.a. & & \gamma_j \geq 0 \;\; j = 1,\ldots,p. \end{array} \tag{38}$$

Therefore, $\hat{y}_i(t)$ is predicted according to the estimated parameters as follows:

$$\hat{y}_i(t) = \hat{\bar{y}}_i + \hat{y}^c_i(t) = \hat{\beta}_0 + \sum_{j=1}^{p}\hat{\beta}_j\bar{x}_{ij} + \sum_{j=1}^{p}\hat{\gamma}_j x^c_{ij}(t). \tag{39}$$

The estimated model explicit the linear relation between a histogram response variable and a set of histogram predictors shared in two components. The first component $\hat{\bar{y}}_i = \hat{\beta}_0 + \sum_{j=1}^{p}\hat{\beta}_j\bar{x}_{ij}$ expresses the linear relation between the mean of the dependent histogram variable and the means of the independent histogram variables. The βs are allowed to be either positive or negative because it is a linear combination of scalar values.
The second component $\hat{y}^c_i(t) = \sum_{j=1}^{p}\hat{\gamma}_j x^c_{ij}(t)$ expresses the relation between qfs apart from their mean value, i.e. the relation among their variability structures, where, with the term variability we intend all the other characteristics of the corresponding HD without considering its mean. In this case, allowing to predict a qf (i.e. a non decreasing function) the γ's cannot be negative.

Goodness of fit (GOF) indices Considering the nature of the data, the evaluation of the goodness of fit (GOF) of the model is not straightforward like for the classic linear regression models. To evaluate the GOF of a regression model on histogram data, we propose to consider the following three indices.

Ω index (Dias and Brito, 2011) The proposed measure is the ratio

$$\Omega = \sum_{i=1}^{n} d_W^2\left(\hat{Y}(i), \bar{y}\right) \Big/ \sum_{i=1}^{n} d_W^2\left(Y(i), \bar{y}\right) \tag{40}$$

(where $\bar{y}$ is the mean of $\bar{Y}$) that varies from 0 to 1 and it is the sum of (Wasserstein) squared distances between quantile functions and a punctual value.

Pseudo $-$ $\mathbf{R^2}$ It is a measure proposed by Verde and Irpino (2010) for the simple linear regression model and here extended to the multivariate case.

In order to build a R^2 index, it is possible to prove (see appendix for details) that SSY can be decomposed as follows:

$$SSY = \sum_{i=1}^{n} d_W^2\left(Y(i), \bar{Y}\right) = \underbrace{\sum_{i=1}^{n} \int_0^1 \left[\hat{y}_i(t) - y_i(t)\right]^2 dt}_{SSE} + \underbrace{\sum_{i=1}^{n} \int_0^1 \left[\bar{y}(t) - \hat{y}_i(t)\right]^2 dt}_{SSR} + \underbrace{-2\left[n \cdot \left(s_{\bar{y}}^2 - \sum_{j=1}^{p} \hat{\gamma}_j \rho(\bar{y}, \bar{x}_j) s_{\bar{y}} s_{\bar{x}_j}\right) + \sum_{j=1}^{p} \hat{\gamma}_j \cdot \frac{\partial SSE}{\partial \gamma_j}(\hat{\gamma}_j)\right]}_{Bias} \tag{41}$$

where the *Bias* term reflects the impossibility of the linear transformation of a set of $\bar{x}(t)$'s of recovering the variability structure of $\bar{y}(t)$.[2] We propose a conservative GOF index considering the following formulation of the $Pseudo - R^2$:

$$PseudoR^2 = \min\left[\max\left[0; 1 - \frac{SSE}{SSY}; \frac{SSR}{SSY}\right]; 1\right]. \tag{42}$$

RMSE The Root Mean Square Error is commonly used as measure of GOF. In our case, consistently with the used distance, we propose the following GOF measure:

$$RMSE_W = \sqrt{\frac{\sum_{i=1}^{n} d_W^2\left(\hat{Y}(i), Y(i)\right)}{n}} = \sqrt{\frac{SSE}{n}}. \tag{43}$$

4 Application on real data

To illustrate the proposed method we present comparison of the existing regression methods for HV on a climatic dataset. The data were obtained from the Clean Air Status and Trends Network (CASTNET)[3], an air quality monitoring network of United States which is designed to provide data to assess trends in air quality, atmospheric deposition, and ecological effects due to changes in air pollutant emissions. In particular, we have chosen to select data about the Ozone concentration in 78 USA sites for which the monitored data was complete.

Ozone is a gas that can cause respiratory diseases. In the literature there exists studies that relates the Ozone concentration level to the Temperature, the Wind speed and the Solar radiation (see for example (Dueñas et al., 2002)). Given the distribution of Temperature (X_1) (Celsius degrees), the distribution of Solar Radiation (X_2) (Watts per square meter) and the

2. Note that $\Gamma \nabla SS(\Gamma)$ depends from the constraints on the solution of the NNLS algorithm for the Γ parameters.
3. http://java.epa.gov/castnet/

	Ozone Concentration (Y in Ppb)	Temperature (X_1 in Celsius deg.)	Solar Radiation (X_2 Watt/m^2)	Wind Speed (X_3 m/s)
Mean (BD)	41.2147	23.2805	645.3507	2.3488
Barycenter mean (VI)	41.2147	23.2805	645.3507	2.3488
Barycenter std (VI)	9.9680	3.7641	225.7818	1.0987
Standard dev. (BD)	13.790	5.3787	252.6736	1.7125
Standard dev. (VI)	9.5295	3.8422	113.4308	1.1337

TAB. 1 – *Ozone dataset: summary statistics. BD is referred to the Billard and Diday (2006) approach, while VI to the Verde and Irpino (2008) one.*

Model	Estimates	Goodness of fit		
		Ω	Ps-R^2	$RMSE$
Billard-Diday	$\hat{y}_i = 18.28 + 0.357\, x_{i1} + 0.017\, x_{i2} + 1.550\, x_{i3}$	0.203	0.024	9.419
Dias-Brito	$\hat{y}_i(t) = 13.32 + 0\, x_{i1}(t) + 0.037\, x_{i2}(t) + 1.691\, x_{i3}(t) +$ $+0\, \tilde{x}_{i1}(t) + 0\, \tilde{x}_{i2}(t) + 0\, \tilde{x}_{i3}(t)$	0.670	0.371	7.557
Irpino-Verde	$\hat{y}_i(t) = 2.93 - 0.346\, \bar{x}_{i1} + 0.07\, \bar{x}_{i2} + 0.395\, \bar{x}_{i3}$ $+0.915\, x^c_{i1}(t) + 0.018\, x^c_{i2}(t) + 1.887\, x^c_{i3}(t)$	0.742	0.460	6.999

TAB. 2 – *Ozone dataset: estimates of the three regression models and GOF indices*

distribution of Wind Speed (X_3) (meters per second), the main objective is to predict the distribution of Ozone Concentration (Y) (Particles per billion) using a linear model. CASTNET collect hourly data and as period of observation we choose the summer seasons of 2010 and the central hours of the days (10 a.m. – 5 p.m.).
For each sites we have collected the histogram data with respect to the four variables.

In table 4 we reported the main summary statistics for the four histogram variables.

Using the full dataset we estimated the three models and the associated goodness of fit indices as reported in table 4.

Observing the good of fitting measures of the three models we can conclude that the proposed and the Dias-Brito model fit better the linear relationship than the Billard-Diday model, and the proposed model is more accurate than the Dias-Brito one. However, the main advantages are related to the interpretation of the models. The Billard-Diday model does not explicit the relationships among the different characteristics of the histogram data. The Dias-Brito model introduce the symmetric distribution concept that can sound particular difficult to explain and to justify. The proposed model allows the researcher to observe the linear relationships among the means in a classic fashion, while it is possible to read the second component in terms of shrinking factor of the variability: for example, we say that an increase of one of Celsius degree in mean will decrease the mean of the Ozone Concentration of 0.346 Ppb, while the variability of the Ozone Concentration increases of about two times (1.877) when the variability of the Wind speed increases of one m/s.

5 Conclusions

The paper present a novel method for linear regression of histogram variables. Considering the nature of the data we proposed to use a particular decomposition of the Wasserstein distance for the definition of the regression model. We showed that the proposed model has in general better interpretation with respect the two main approaches presented in the literature. We consider to address new efforts in the direction of investigate the properties of the involved estimators and in particular of the bias term.

References

Afonso, F., L. Billard, E. Diday, and M. Limam (2008). Symbolic linear regression methodology. In *Symbolic Data Analysis and the SODAS Software*, pp. 359–372. Wiley.

Bertrand, P. and F. Goupil (2000). Descriptive statistics for symbolic data. In H.-H. Bock and E. Diday (Eds.), *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*, pp. 103–124. Springer Berlin Heidelberg.

Billard, L. and E. Diday (2006). *Symbolic data analysis: conceptual statistics and data mining*. Wiley.

Bock, H. and E. Diday (2000). *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*. Springer verlag.

Cuesta-Albertos, J. A., C. Matrán, and A. Tuero-Díaz (1997). Optimal transportation plans and convergence in distribution. *J. Multivar. Anal. 60*, 72–83.

Dias, S. and P. Brito (2011). A new linear regression model for histogram-valued variables. In *58th ISI World Statistics Congress*, Dublin, Ireland.

Diday, E. and M. Noirhomme-Fraiture (2008). *Symbolic Data Analysis and the SODAS software*. Wiley.

Dueñas, C., M. Fernández, S. Cañete, J. Carretero, and E. Liger (2002). Assessment of ozone variations and meteorological effects in an urban area in the mediterranean coast. *Science of The Total Environment 299*(1-3), 97 – 113.

Irpino, A. and E. Romano (2007). Optimal histogram representation of large data sets: Fisher vs piecewise linear approximation. *RNTI E*(9), 99–110.

Irpino, A., R. Verde, and Y. Lechevallier (2006). Dynamic clustering of histograms using wasserstein metric. In *COMPSTAT 2006*, pp. 869–876. Physica-Verlag.

Lawson, C. L. and R. J. Hanson (1974). *Solving Least Square Problems*. Edgeworth Cliff, NJ: Prentice Hall.

Lima Neto, E. d. A. and F. d. A. T. de Carvalho (2010). Constrained linear regression models for symbolic interval-valued variables. *Comput. Stat. & Data Analysis 54*(2), 333–347.

Pearson, K. (1895). Contributions to the mathematical theory of evolution. ii. skew variation in homogeneous material. *Phil Trans R Soc A* (186), 343 – 414.

Rüschendorf, L. (2001). Wasserstein metric. In M. Hazewinkel (Ed.), *Encyclopedia of Mathematics*. Springer.

Verde, R. and A. Irpino (2007). Dynamic clustering of histogram data: Using the right metric. In P. Brito, G. Cucumel, P. Bertrand, and F. de Carvalho (Eds.), *Selected Contributions in Data Analysis and Classification*, pp. 123–134. Springer Berlin Heidelberg.

Verde, R. and A. Irpino (2008). Comparing histogram data using a mahalanobis-wasserstein distance. In P. Brito (Ed.), *COMPSTAT 2008*, Chapter 7, pp. 77–89. Physica-Verlag HD.

Verde, R. and A. Irpino (2010). Ordinary least squares for histogram data based on wasserstein distance. In Y. Lechevallier and G. Saporta (Eds.), *Proceedings of COMPSTAT'2010*, Chapter 60, pp. 581–588. Heidelberg: Physica-Verlag HD.

APPENDIX The decomposition of the sum of squares of Y

Considering $\mathbf{1} = [1]_{n\times 1}$, SSY can be written as:

$$SSY = n \cdot s_y^2 = \sum_{i=1}^{n} d_W^2\left(y_i(t), \bar{y}(t)\right) = \sum_{i=1}^{n} \int_0^1 \left[y_i(t) - \bar{y}(t)\right]^2 dt$$

further

$$SSY = \sum_{i=1}^{n} \int_0^1 \left[y_i(t) - \bar{y}(t) + \hat{y}_i(t) + \hat{y}_i(t)\right]^2 dt =$$
$$= \underbrace{\sum_{i=1}^{n} \int_0^1 e_i^2(t)dt}_{SSE} + \underbrace{\sum_{i=1}^{n} \int_0^1 \left(\hat{y}_i(t) - \bar{y}(t)\right)^2 dt}_{SSR} - 2\sum_{i=1}^{n} \int_0^1 \left(\bar{y}(t) - \hat{y}_i(t)\right) e_i(t)dt$$

Differently that in the OLS estimate regression model the term:

$$Bias = \sum_{i=1}^{n} \int_0^1 \left(\bar{y}(t) - \hat{y}_i(t)\right) e_i(t)dt \neq 0$$

It can be write as:

$$Bias = \underbrace{\sum_{i=1}^{n} \int_0^1 \bar{y}_i(t)e_i(t)dt}_{I} - \underbrace{\sum_{i=1}^{n} \int_0^1 \hat{y}_i(t)e_i(t)dt}_{II}$$

For (I), replacing $e_i(t) = y_i(t) - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j \bar{x}_{ij} - \sum_{j=1}^{p} \hat{\gamma}_j x_{ij}^c(t)$ we have:

$$\sum_{i=1}^{n}\int_0^1 \bar{y}_i(t)e_i(t)dt = \int_0^1 \bar{y}(t)\sum_{i=1}^{n} y_i(t)dt - \int_0^1 \bar{y}(t)dt \cdot \sum_{i=1}^{n} \underbrace{(\hat{\beta}_0 + \sum_{j=1}^{p}\hat{\beta}_j\bar{x}_{ij})}_{=\bar{y}_i} +$$

$$- \sum_{i=1}^{n}\sum_{j=1}^{p}\hat{\gamma}_j \int_0^1 \bar{y}(t)x_{ij}^c(t)dt =$$

$$= n\cdot\sigma_{\bar{y}}^2 + n\bar{y}^2 - n\bar{y}^2 - n\sum_{j=1}^{p}\hat{\gamma}_p r_{\bar{y}\bar{x}_j}\sigma_{\bar{y}}\sigma_{\bar{x}_j} = n\cdot\sigma_{\bar{y}}^2 - n\sum_{j=1}^{p}\hat{\gamma}_p r_{\bar{y}\bar{x}_j}\sigma_{\bar{y}}\sigma_{\bar{x}_j} =$$

$$= n\cdot\left(\sigma_{\bar{y}}^2 - \sum_{j=1}^{p}\hat{\gamma}_p r_{\bar{y}\bar{x}_j}\sigma_{\bar{y}}\sigma_{\bar{x}_j}\right)$$

(note that: $\sigma_{\bar{x}_j^c} = \sigma_{\bar{x}_j}$ and $r_{\bar{y}\bar{x}_j^c} = r_{\bar{y}\bar{x}_j}$)
For (II), being

$$\hat{y}_i(t) = \hat{\beta}_0 + \sum_{j=1}^{p}\hat{\beta}_j\bar{x}_{ij} + \sum_{j=1}^{p}\hat{\gamma}_j x_{ij}^c(t)$$

and

$$e_i(t) = y_i(t) - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j\bar{x}_{ij} - \sum_{j=1}^{p}\hat{\gamma}_j x_{ij}^c(t)$$

and indicating with $\sum_{j=0}^{p}\hat{\beta}_j\bar{x}_{ij}^+ = \hat{\beta}_0 + \sum_{j=1}^{p}\hat{\beta}_j\bar{x}_{ij}$ (where $\bar{x}_{i0}^+ = 1$), we have:

$$\sum_{i=1}^{n}\int_0^1 \hat{y}_i(t)e_i(t)dt =$$

$$= \sum_{i=1}^{n}\int_0^1 \left(\sum_{j=0}^{p}\hat{\beta}_j\bar{x}_{ij}^+ + \sum_{j=1}^{p}\hat{\gamma}_j x_{ij}^c(t)\right)\left(y_i(t) - \sum_{j=0}^{p}\hat{\beta}_j\bar{x}_{ij}^+ - \sum_{j=1}^{p}\hat{\gamma}_j x_{ij}^c(t)\right)dt =$$

$$= \sum_{i=1}^{n}\int_0^1 \left[\sum_{j=0}^{p}\hat{\beta}_j\bar{x}_{ij}^+ y_i(t) - \left(\sum_{j=0}^{p}\hat{\beta}_j\bar{x}_{ij}^+\right)^2 - \left(\sum_{j=0}^{p}\hat{\beta}_j\bar{x}_{ij}^+\right)\left(\sum_{j=1}^{p}\hat{\gamma}_j x_{ij}^c(t)\right) + \right.$$

$$\left. + \sum_{j=1}^{p}\hat{\gamma}_j x_{ij}^c(t)y_i(t) - \left(\sum_{j=1}^{p}\hat{\gamma}_j x_{ij}^c(t)\right)\left(\sum_{j=0}^{p}\hat{\beta}_j\bar{x}_{ij}^+\right) - \left(\sum_{j=1}^{p}\hat{\gamma}_j x_{ij}^c(t)\right)^2\right]dt =$$

$$= \sum_{i=1}^{n}\sum_{j=0}^{p}\hat{\beta}_j\bar{x}_{ij}^+\int_0^1 y_i(t)dt - \sum_{i=1}^{n}\left(\sum_{j=0}^{p}\hat{\beta}_j\bar{x}_j^+\right)^2 - 2\sum_{i=1}^{n}\left(\sum_{j=0}^{p}\hat{\beta}_j\bar{x}_j^+\right)\sum_{j=1}^{p}\hat{\gamma}_j\underbrace{\int_0^1 x_{ij}^c(t)dt}_{=0} +$$

$$+ \sum_{i=1}^{n}\sum_{j=1}^{p}\hat{\gamma}_j\int_0^1 x_{ij}^c(t)y_i(t)dt - \sum_{i=1}^{n}\int_0^1\sum_{j=1}^{p}\hat{\gamma}_j x_{ij}^c(t)^2\,dt =$$

$$= \sum_{j=0}^{p}\hat{\beta}_j\sum_{i=1}^{n}\bar{x}_j^+\bar{y}_i - \sum_{i=0}^{p}\sum_{j'=0}^{p}\sum_{i=1}^{n}\hat{\beta}_j\hat{\beta}_{j'}\bar{x}_{ij}^+\bar{x}_{ij'}^+ +$$

$$+ \sum_{i=1}^{n}\sum_{j=1}^{p}\hat{\gamma}_j\int_0^1 x_{ij}^c(t)y_i(t)dt - \sum_{i=1}^{n}\int_0^1\left(\sum_{j=1}^{p}\sum_{j'=1}^{p}\hat{\gamma}_j\hat{\gamma}_{j'}x_{ij}^c(t)x_{ij'}^c(t)\right)dt =$$

$$= \sum_{j=0}^{p} \hat{\beta}_j \underbrace{\left[\sum_{i=1}^{n} \bar{x}_{ij}^{+} \left(\bar{y}_i - \sum_{j'=0}^{p} \hat{\beta}_{j'} \bar{x}_{ij'}^{+} \right) \right]}_{C1=0} +$$

$$+ \sum_{j=1}^{p} \hat{\gamma}_j \underbrace{\left[\sum_{i=1}^{n} \int_0^1 x_{ij}^c(t) y_i(t) dt - \sum_{j=1}^{p} \hat{\gamma}_{j'} \int_0^1 x_{ij}^c(t) x_{ij'}^c(t) dt \right]}_{C2}$$

The expression in $C1$ is equal to 0 according to the first order conditions for each $\hat{\beta}_j$ in the OLS estimation solution: $\frac{\partial SSE}{\partial \beta_j}(\hat{\beta}_j) = 0$. The expression in $C2$ is also related to the first order condition for each $\hat{\gamma}_j$, but it could be not equal to 0 because of the constraints for the solutions of the Non Negative Least Squared algorithm.
Then, denoting

$$C2 = \frac{\partial SSE}{\partial \gamma_j}(\hat{\gamma}_j)$$

we can rewrite

$$\sum_{j=1}^{p} \hat{\gamma}_j \left[\sum_{i=1}^{n} \left(\int_0^1 x_{ij}^c(t) y_i(t) dt - \sum_{j=1}^{p} \hat{\gamma}_{j'} \int_0^1 x_{ij}^c(t) x_{ij'}^c(t) dt \right) \right] = \sum_{j=1}^{p} \hat{\gamma}_j \cdot \frac{\partial SSE}{\partial \gamma_j}(\hat{\gamma}_j).$$

The $bias$ term is

$$bias = -2 \left[n \cdot \left(\sigma_{\bar{y}}^2 - \sum_{j=1}^{p} \hat{\gamma}_j r_{\bar{y}\bar{x}_j} \sigma_{\bar{y}} \sigma_{\bar{x}_j} \right) + \sum_{j=1}^{p} \hat{\gamma}_j \cdot \frac{\partial SSE}{\partial \gamma_j}(\hat{\gamma}_j) \right]$$

and, finally the SSY decomposition is:

$$SSY = SSE + SSR - 2 \left[n \cdot \left(\sigma_{\bar{y}}^2 - \sum_{j=1}^{p} \hat{\gamma}_j r_{\bar{y}\bar{x}_j} \sigma_{\bar{y}} \sigma_{\bar{x}_j} \right) + \sum_{j=1}^{p} \hat{\gamma}_j \cdot \frac{\partial SSE}{\partial \gamma_j}(\hat{\gamma}_j) \right].$$