Ordinary Least Squares for Histogram Data based on Wasserstein Distance

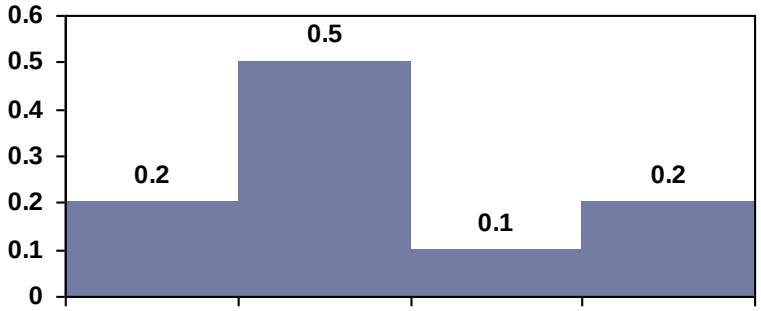
Rosanna Verde Antonio Irpino
Dipartimento di Studi Europei e Mediterranei
Seconda Università degli Studi di Napoli (ITALY)
[rosanna.verde] [antonio.irpino]@unina2.it

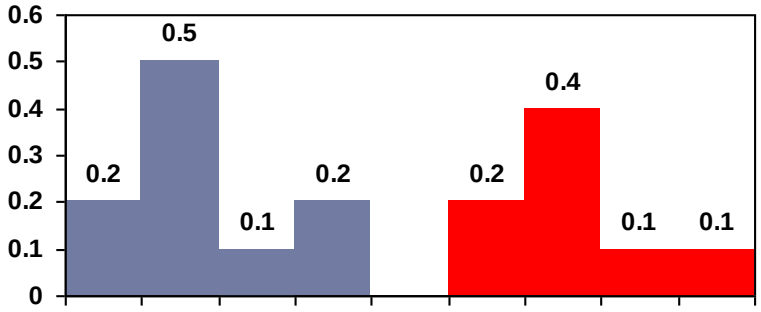
Outline

- Histogram data
- A regression model for histogram variables
- Properties of the Wasserstein distance
- Ordinary Least Square fitting
- Tools for the interpretation
- An application on real data

Sources of histogram data

- Result of summary/clustering procedures
  - From surveys
  - From large databases
  - From sensors
    - Temperatures
    - Pollutant concentration
    - Network activity
- Data streams
  - Description of time windows
- Image analysis
  - Color bandwidths
- Confidentiality data
  - Summary data – non punctual





Histogram data as a particular case of modal symbolic descriptions [*Bock and Diday (2000)*]

Histogram data is a kind of symbolic representation which allows to describe an individual by means of a histogram

In *Bock and Diday (2000)* **Histogram variable** is one of the three definition of modal numerical variables :

- [**Histogram variable**] The description is a classic histogram where the support is partitioned into intervals. Each interval is weighted by the empirical density;
- [**Empirical distribution function variable**] The description is done according to an empirical distribution function;
- [**Model of distribution variable**] The description is done according to a predefined model of random variable.

Histogram variable

Let Y be a continuous variable defined on a finite support $\mathbf{S} = [\underline{y}; \overline{y}]$, where: $\underline{y}$ and $\overline{y}$ are the minimum and maximum values of the variable domain.

The variable Y is partitioned into a set of contiguous intervals (bins)

$$\{I_1, \ldots, I_h, \ldots, I_H\}, \text{ where } I_h = [\underline{y}_h; \overline{y}_h).$$

Given *n* observations of the variable Y, each semi-open interval, I_h is associated with a random variable equal to

$$\Psi(I_h) = \sum_{u=1}^{N} \Psi_{y_u}(I_h) \text{ where } \Psi_{y_u}(I_h) = 1 \text{ if } y_u \in I_h \text{ and } 0 \text{ otherwise.}$$

It is possible to associate with I_h an empirical distribution $\pi_h = \Psi(I_h)/N.$

A histogram of Y is the representation in which each pair (I_h, π_h) (for $h = 1, \ldots, H$) is represented by a vertical bar, with base interval I_h along the horizontal axis and the area proportional to π_h.

A Regression model for histogram variables

- In order to study the dependence structure of a histogram variable Y (dependent) to the another X (independent)

  we introduce a new regression approach based on the Ordinary Least Square estimation method

  According to the nature of the variables, we propose to compute the squared deviations (in the least squares function) by using the Wasserstein distance.

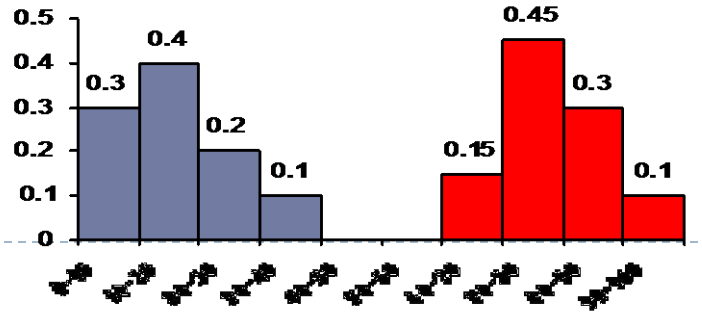
A Regression model for histogram variables

Data = Model Fit + Residual

- Linear regression is a general method for estimating/describing association between a **continuous** outcome variable (dependent) and one or multiple predictors in one equation.

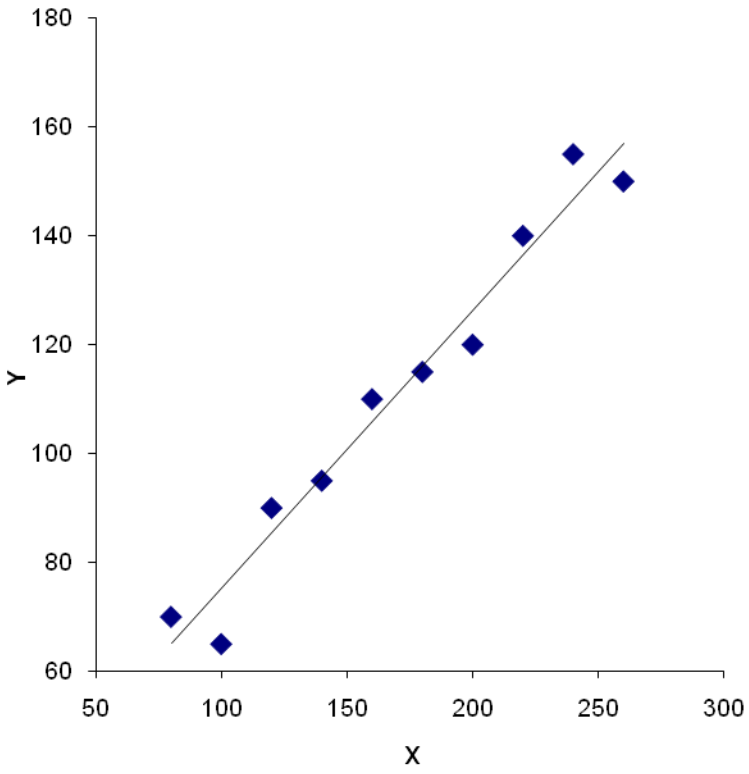
Easy conceptual task with classic data

But what does it means when dealing with histogram data?

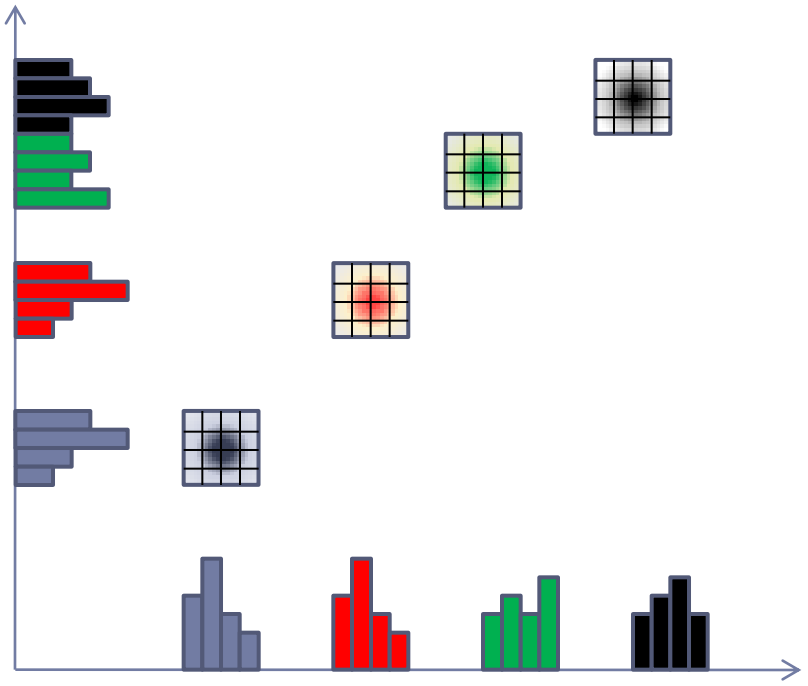


Simple linear regression

Classic data



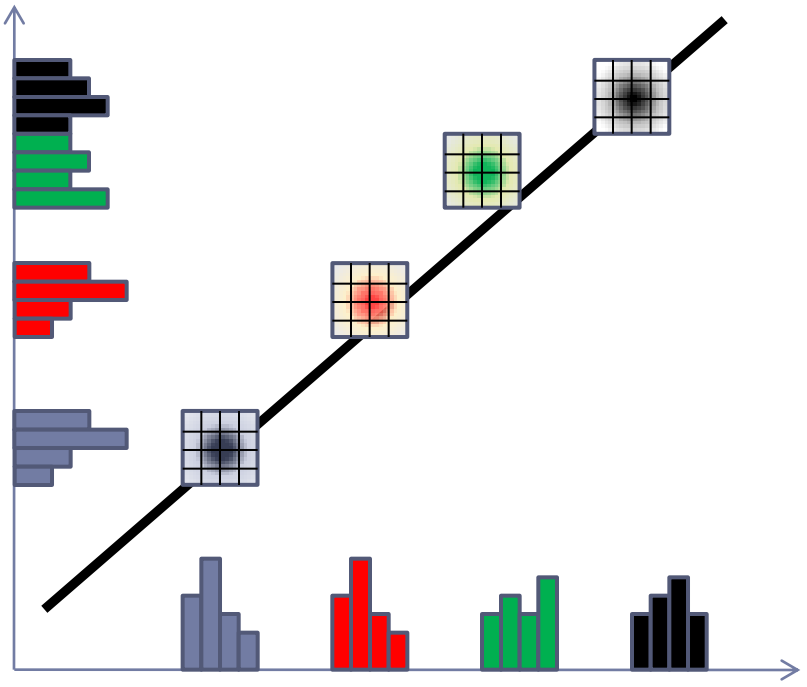
Histogram data



Regression between histograms: a proposal

A solution was given by Billard and Diday (2006)

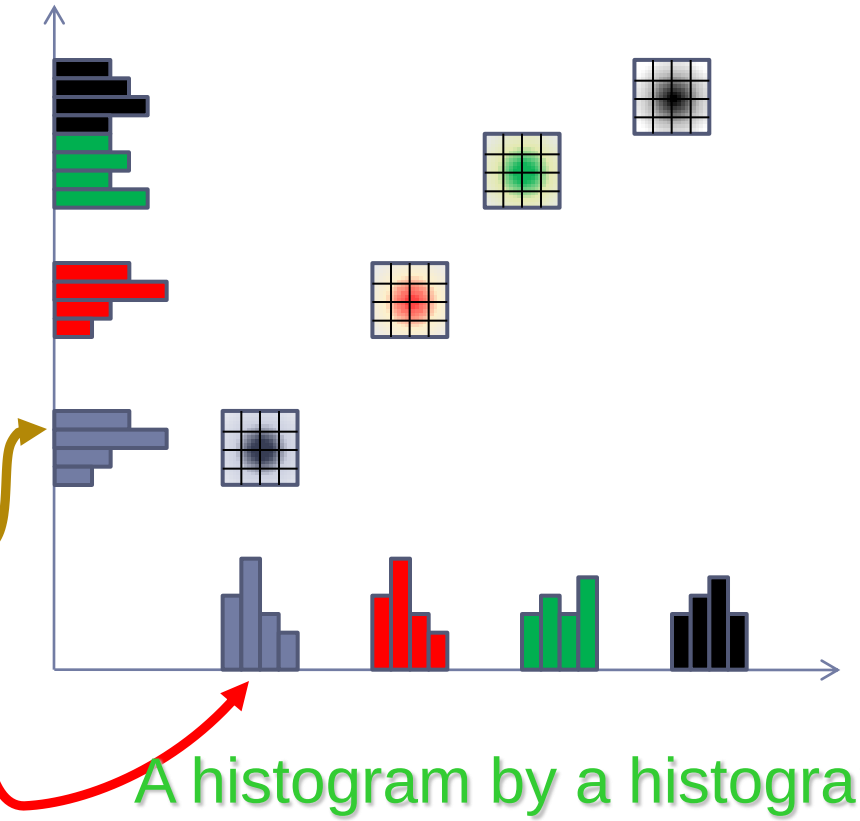
- The model fit a linear regression line throught the mixture of the n bivariate distributions
- Given a punctual value of X it is possible to predict the punctual value of Y



Regression between histograms: our approach

- Given a histogram description for X, we search for **a linear trasformation** of the description which allows us to predict the histogram description of Y

- For example:
  given **the temperature histogram observed in a region during a month**,
  Is it possible to predict **the distribution of the temperature of another month** using a linear transformation of the histogram variable?



Wasserstein distance

- We propose to use the Wasserstein-Kantorovich metric in Least Square Function.

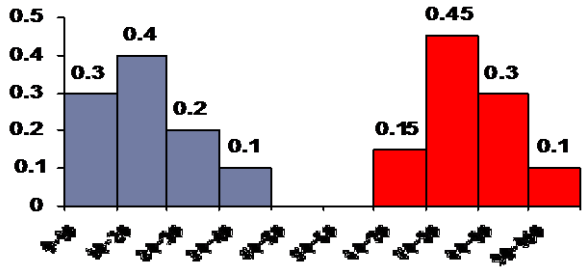
  Expecially the derived L_2-Mallow's distance between two quantile functions

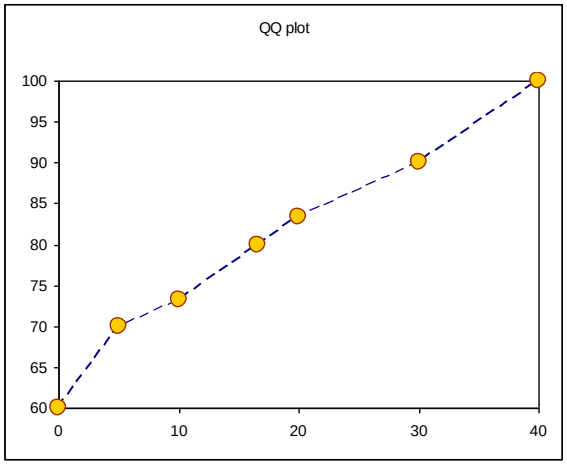
$$d_W\left(x_i, x_j\right) = \sqrt{\int_0^1 \left(F_i^{-1}(t) - F_j^{-1}(t)\right)^2 dt}$$

An interpretative decomposition of the L_2-Wasserstein metric

$$d_W^2(x_i, x_j) := \int_0^1 \left(F_i^{-1}(t) - F_j^{-1}(t) \right)^2 dt =$$

$$= \underbrace{\left(\bar{x}_i - \bar{x}_j \right)^2}_{Location} + \underbrace{\left(\sigma_i - \sigma_j \right)^2}_{Size} + \underbrace{2\sigma_i\sigma_j(1 - \rho(x_i, x_j))}_{Shape}$$





If the two distribution have the same shape:

$$d_W^2(x_i, x_j) := \underbrace{\left(\mu_i - \mu_j \right)^2}_{Location} + \underbrace{\left(\sigma_i - \sigma_j \right)^2}_{Size}$$

If they have the same size and shape: $d_W^2(x_i, x_j) := \underbrace{\left(\mu_i - \mu_j \right)^2}_{Location}$

Some simplifications and notations

Mean and variance of the distribution/ histogram data

$$x_i(t) = F_i^{-1}(t)$$

quantile function of the *i-th* macro-unit x_i (histogram/distribution data)

$$\bar{x}_i = \int_0^1 x_i(t)dt \text{ and } \sigma_{x_i}^2 = \int_0^1 [x_i(t)]^2 dt - [\bar{x}_i]^2 \Rightarrow \int_0^1 [x_i(t)]^2 dt = \sigma_{x_i}^2 + [\bar{x}_i]^2$$

Average distribution/ histogram data

$$\bar{x}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t) \quad \forall t \in [0,1]; \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 x_i(t)dt = \frac{1}{n}\sum_{i=1}^{n}\bar{x}_i = \int_0^1 \bar{x}(t)dt$$

$$\rho(x_i, x_j) = \frac{\int_0^1 x_i(t)x_j(t)dt - \bar{x}_i\bar{x}_j}{\sigma_{x_i}\sigma_{x_j}} \Rightarrow \int_0^1 x_i(t)x_j(t)dt = \rho(x_i, x_j)\sigma_{x_i}\sigma_{x_j} + \bar{x}_i\bar{x}_j$$

Correlation between pair of distribution/histogram data (x_i , x_j)

Fitting with a linear model

- Given two variables Y and X regression model is here proposed to perform a linear transformation of X which better fit Y

$$y_i(t) = \underbrace{\alpha + \beta x_i(t)}_{\hat{y}_i} + \varepsilon_i(t) \quad \forall t \in [0,1]$$

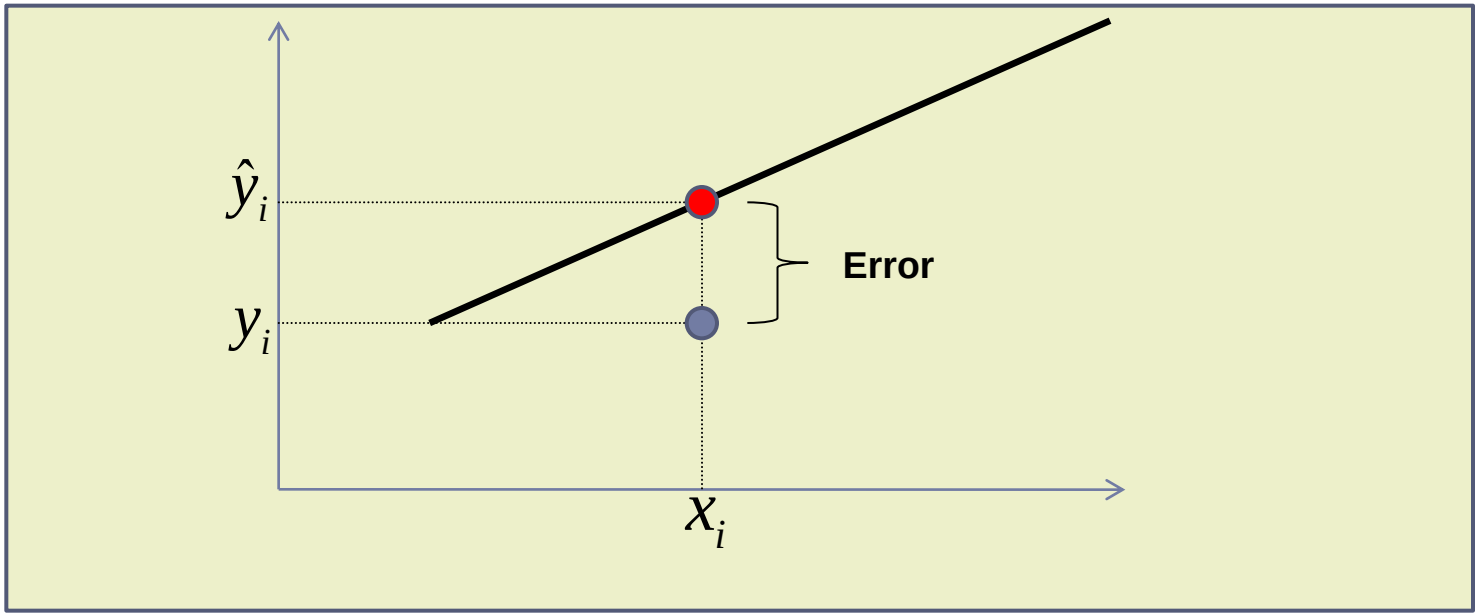
- Considering the error as close as possible to zero:

$$\varepsilon_i(t) = y_i(t) - \hat{y}_i(t)$$

The error term in the classic case

- Classic case (Euclidean norm)

$$\varepsilon_i = y_i - \hat{y}_i \qquad \varepsilon_i^2 = \left(y_i - \hat{y}_i\right)^2 = d_E^2\left(y_i, \hat{y}_i\right)$$

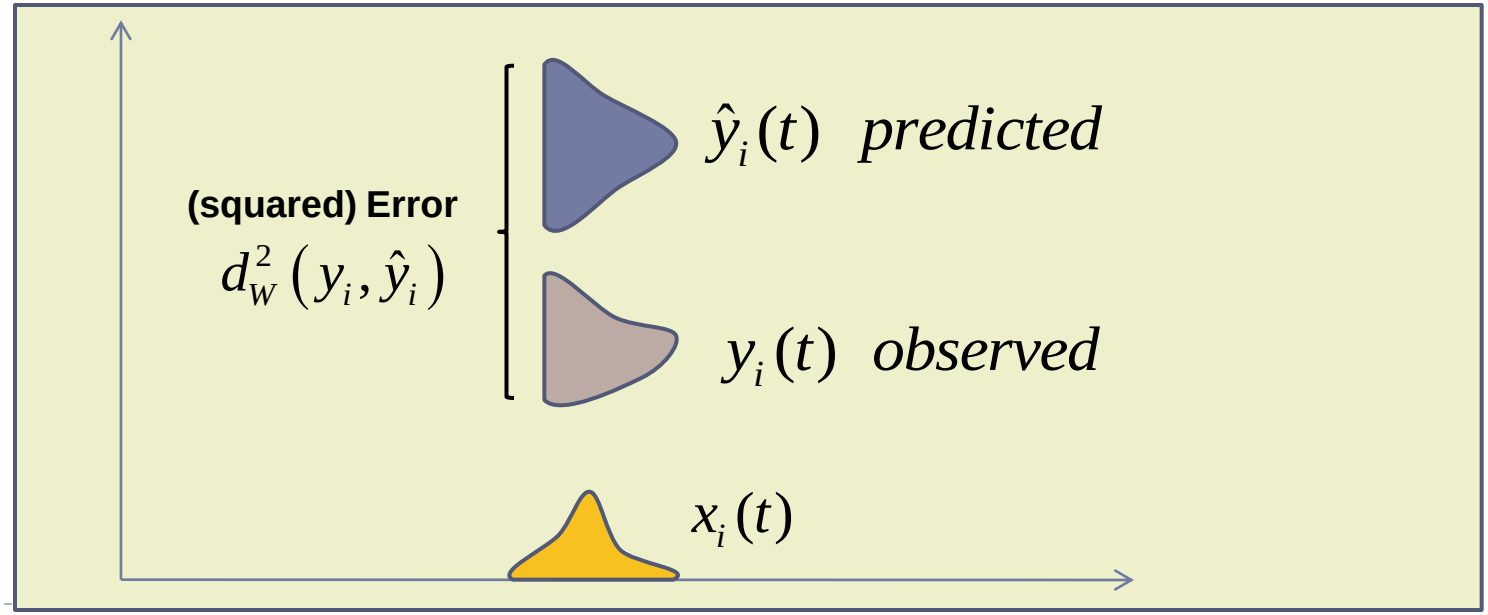


The error term of the model (our approach)

- Histogram case (Wasserstein distance)

$$\varepsilon_i(t) = y_i(t) - \hat{y}_i(t) \quad \left[\varepsilon_i(t)\right]^2 = \left[y_i(t) - \hat{y}_i(t)\right]^2 \ \forall t \in [0,1]$$

$$\int_0^1 \left[y_i(t) - \hat{y}_i(t)\right]^2 dt = d_W^2\left(y_i, \hat{y}_i\right)$$



Fitting a linear model: histograms

- We propose to find a linear transformation of the quantile function of x_i (histogram data) in order to predict the quantile function of y_i i.e.:

$$\hat{y}_i(t) = f(x_i(t)) = \alpha + \beta x_i(t) \quad \forall t \in [0,1]$$

It is worth noting the linear transformation is unique: the parameters α and β are estimated for all the i macro-units x_i and y_i

- A first problem:

  Only if $\beta>0$ a quantile function $\hat{y}_i(t)$ can be derived.

  In order to overcome this problem, we propose a solution based on the decomposition of the Wasserstein distance.

Solution to $\beta<0$

- The quantile function can be decomposed as:

$$x_i(t) = \overline{x}_i + x_i^c(t) \quad where$$

$$x_i^c(t) = x_i(t) - \overline{x}_i \quad is\ the\ centered\ quantile\ function$$

- Then, we propose the following model:

$$y_i(t) = \alpha + \beta_1 \overline{x}_i + \beta_2 x_i^c(t) + \varepsilon_i(t).$$

- Using the Wasserstein distance it is possible to set up a OLS method that returns three coefficients.

  We demonstrate β_2 is always greater or equal to zero.

The error term:

a property of the Wasserstein distance decomposition

- The (squared) error can be written according the two components

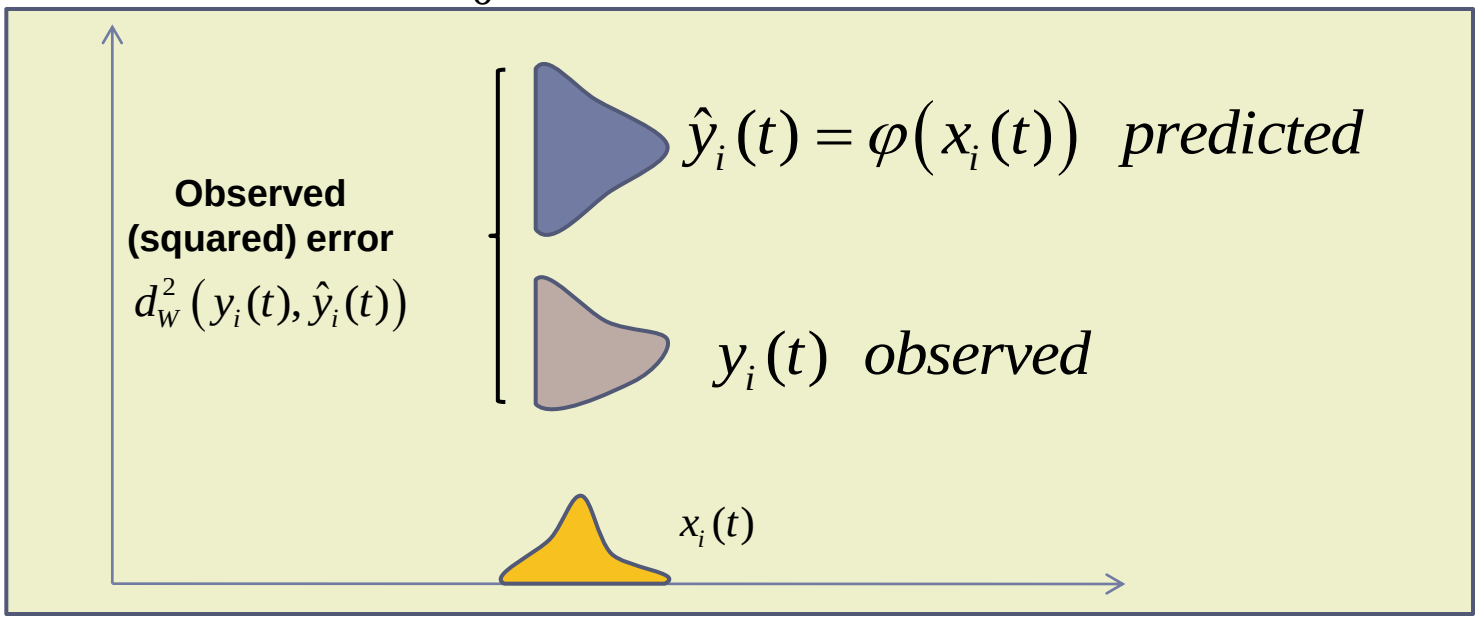
$$\varepsilon_i^2 = d_W^2(y_i, \hat{y}_i) = \int_0^1 \left(y_i(t) - \hat{y}_i(t)\right)^2 dt =$$

$$= \left(\bar{y}_i - \hat{\bar{y}}_i\right)^2 + d_W^2(y_i^c, \hat{y}_i^c)$$

Ordinary Least Squares

$$\underset{(\alpha,\beta_1,\beta_2)\in\Re^3}{\arg\min}\ f(\alpha,\beta_1,\beta_2)=\sum_{i=1}^{n}\varepsilon_i^{\,2}(t)=\sum_{i=1}^{n}d_W^2\left(y_i(t),\hat{y}_i(t)\right)$$

$$f(\alpha,\beta_1,\beta_2)=\sum_{i=1}^{n}\int_0^1\left[y_i(t)-\alpha-\beta_1\overline{x}_i-\beta_2 x_i^c(t)\right]^2 dt.$$



Solving OLS

- First order conditions

$$\begin{cases} \dfrac{\delta f}{\delta \alpha} = -2\sum_{i=1}^{n}\int_{0}^{1}\left(\bar{y}_i + y_i^c(t) - \alpha - \beta_1 \bar{x}_i - \beta_2 x_i^c(t)\right)dt = 0 & (I) \\ \dfrac{\delta f}{\delta \beta_1} = -2\sum_{i=1}^{n}\int_{0}^{1}\bar{x}_i\left(\bar{y}_i + y_i^c(t) - \alpha - \beta_1 \bar{x}_i - \beta_2 x_i^c(t)\right)dt = 0 & (II) \\ \dfrac{\delta f}{\delta \beta_2} = -2\sum_{i=1}^{n}\int_{0}^{1}x_i^c(t)\left(\bar{y}_i + y_i^c(t) - \alpha - \beta_1 \bar{x}_i - \beta_2 x_i^c(t)\right)dt = 0 & (III) \end{cases}$$

The estimated parameters

Correlation between quantile functions $x_i\ y_i$

$$\hat{\alpha} = \bar{y} - \hat{\beta}_1 \bar{x}; \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} \bar{x}_i \bar{y}_i - n\bar{y}\,\bar{x}}{\sum_{i=1}^{n} \bar{x}_i^2 - n\bar{x}^2}; \quad \hat{\beta}_2 = \frac{\sum_{i=1}^{n} \rho(x_i, y_i)\sigma_{x_i}\sigma_{y_i}}{\sum_{i=1}^{n} \sigma_{x_i}^2}$$

- It is easy to see that:

$$\hat{\alpha}, \hat{\beta}_1 \in \Re \ \ and \ \ \hat{\beta}_2 \geq 0$$

Interpretation of the parameters

- Regression parameters for the distribution mean locations

$$\hat{\alpha}, \hat{\beta}_1 \in \Re$$

- Shrinking factor for the variability

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} \rho(x_i, y_i)\sigma_{x_i}\sigma_{y_i}}{\sum_{i=1}^{n} \sigma_{x_i}^2} \geq 0$$

- >1 (<1) the y_i histogram has a greater (smaller) variability than the x_i histogram.
- =0 when the distributions collapse in points.

Tools for the interpretation

- The sum of squares of Y is

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

- We recall the decomposition of the Sum of Squares of Y

$$\mathrm{SS(Y)} = \mathrm{SS}_{Error} + \mathrm{SS}_{Regression}$$

Decomposition of SS(Y)

- Being: $\hat{y}_i(t) = \hat{\alpha} + \hat{\beta}_1 \bar{x}_i + \hat{\beta}_2 x_i^c(t)$

- we obtain

$$SS(Y) = \sum_{i=1}^{n} d_W^2\left(y_i(t), \bar{y}(t)\right) = \underbrace{\sum_{i=1}^{n} \int_0^1 \left[\hat{y}_i(t) - y_i(t)\right]^2 dt}_{SS_{Error}} +$$

$$+ \underbrace{\sum_{i=1}^{n} \int_0^1 \left[\bar{y}(t) - \hat{y}_i(t)\right]^2 dt}_{SS_{Regression}} \underbrace{-2n \int_0^1 \bar{y}(t)\bar{e}(t)dt}_{Bias}$$

$$\bar{e}(t) = \frac{1}{n} \sum_{i=1}^{n} (y_i(t) - \hat{y}_i(t)) \quad \forall t \in [0,1]$$

Average error function

The bias

- The bias is due to different shapes of distributions:

$$bias = \int_0^1 \bar{y}(t)\bar{e}(t)dt = \sigma_{\bar{y}}^2 - \beta_2 \rho(\bar{x}(t), \bar{y}(t))\sigma_{\bar{x}}\sigma_{\bar{y}}$$

Correlation between the average quantile functions

- $bias=0$ when all the histograms have the same shape

That represents the incapacity of the linear transformation of fitting distributions that are very different in shape

A measure of fitting

- Pseudo R^2

  Considering that

$$SS_{Regression} = \sum_{i=1}^{n}\left(\bar{y}_i - \bar{\hat{y}}_i\right)^2 + \sum_{i=1}^{n}\int_0^1 \left[\bar{y}^c(t) - \hat{y}_i^c(t)\right]^2 dt - \int_0^1 \bar{y}(t)\bar{e}(t)dt$$

  We propose the following pseudo R^2

$$PseudoR^2 = \min\left[\max\left[0;1-\frac{SS_{Error}}{SS(Y)}\right];1\right].$$

An application on a Climatic Dataset: 60 Chinese stations

Histogram data

Let us be considered to predict the following fenomena:

- Humidity
- Pressure
- Temperature
- Wind Speed
- Precipitation

in July from the distributions observed in January

♯	Variable	μ_j	σ_j	$VAR_F(X_j)$	$STD_F(X_j)$
X_1	Mean Relative Humidity (percent) Jan	67.9	7.0	127.9	11.3
X_2	Mean Relative Humidity (percent) July	73.9	4.5	114.2	10.7
X_3	Mean Station Pressure(mb) Jan	968.3	3.6	5864.7	76.5
X_4	Mean Station Pressure(mb) July	951.1	3.0	5084.4	71.3
X_5	Mean Temperature (Cel.) Jan	-1.2	1.7	114.8	10.7
X_6	Mean Temperature (Cel.) July	25.2	1.0	11.3	3.4
X_7	Mean Wind Speed (m/s) Jan	2.3	0.6	1.1	1.0
X_8	Mean Wind Speed (m/s) July	2.3	0.5	0.6	0.8
X_9	Total Precipitation (mm) Jan	18.2	14.3	519.6	22.7
X_{10}	Total Precipitation (mm) July	144.6	80.8	499.9	70.7

Main Results

Variable	Y	X	α	β_1	β_2	PseudoR²	Bias/SS(Y)
Relative Umidity (%)	July	January	472.52	0.393	0.593	0.1564	-0.0296
Station Pressure (mb)	July	January	515.31	0.929	0.993	**0.9981**	0.0007
Temperature (Cel)	July	January	25.46	0.196	0.521	0.3813	-0.0185
Wind Speed (m/s)	July	January	7.98	0.638	0.848	0.6563	-0.0564
Precipitation (mm)	July	January	1337.22	0.617	3.578	**0.0000**	-0.9275

Best fitting model: Station pressure July ← January

β_2 shows as the distribution have quite the same variability while the *bias value* puts in evidence that the histograms have quite the same shape

Variable	Y	X	α	β_1	β_2	PseudoR²	Bias/SS(Y)
Relative Umidity (%)	July	January	472.52	0.393	0.593	0.1564	-0.0296
Station Pressure (mb)	July	January	515.31	0.929	0.993	**0.9981**	0.0007
Temperature (Cel)	July	January	25.46	0.196	0.521	0.3813	-0.0185
Wind Speed (m/s)	July	January	7.98	0.638	**0.848**	0.6563	-0.0564
Precipitation (mm)	July	January	1337.22	0.617	3.578	**0.0000**	-0.9275

The estimated parameter β_2 in the model Wind speed July ← January

shows that the variability of the predicted distribution on July is smaller than the January one

Variable	Y	X	α	β_1	β_2	PseudoR²	Bias/SS(Y)
Relative Umidity (%)	July	January	472.52	0.393	0.593	0.1564	-0.0296
Station Pressure (mb)	July	January	515.31	0.929	0.993	**0.9981**	0.0007
Temperature (Cel)	July	January	25.46	0.196	0.521	0.3813	-0.0185
Wind Speed (m/s)	July	January	7.98	0.638	0.848	0.6563	-0.0564
Precipitation (mm)	July	January	1337.22	0.617	3.578	**0.0000**	-0.9275

The worst fitting model: Precipitation July ← January

The January variability is explained only by the different shape components

The *PseudooR*^{2} and *bias* values show has the histogram data present very different shapes

That makes unable a linear model to explain the causal relationship between this Histogram variables

Main references

- BILLARD, L. and DIDAY, E. (2006): Symbolic Data Analysis: Conceptual Statistics and Data Mining. Wiley Series in Computational Statistics. John Wiley & Sons.
- BOCK, H.H. and DIDAY, E. (2000): Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data. Studies in Classification, Data Analysis and Knowledge Organisation, Springer-Verlag.
- CUESTA-ALBERTOS, J.A., MATRAN, C., TUERO-DIAZ, A. (1997): Optimal transportation plans and convergence in distribution. Journ. of Multiv. An., 60, 72–83.
- GIBBS, A.L. and SU, F.E. (2002): On choosing and bounding probability metrics. Intl. Stat. Rev. 7 (3), 419–435.
- IRPINO, A., LECHEVALLIER, Y. and VERDE, R. (2006): Dynamic clustering of histograms using Wasserstein metric. In: Rizzi, A., Vichi, M. (eds.) COMPSTAT 2006. Physica-Verlag, Berlin, 869–876.
- VERDE, R. and IRPINO, A.(2008): Comparing Histogram data using a Mahalanobis–Wasserstein distance. In: Brito, P. (eds.) COMPSTAT 2008. Physica–Verlag, Springer, Berlin, 77–89.
- LIMA NETO, E.d.A. and DE CARVALHO, F.d.A.T. (2010): Constrained linear regression models for symbolic interval-valued variables. Computational Statistics and Data Analysis, 54, 2, Elsevier, 333–347.