Automatic Discovery of the Statistical Types of Variables in a Dataset

Alla Usova a.usova@studenti.unipi.it
Steffania Sierra Galvis s.sierragalvis@studenti.unipi.it

IN SUPREMÆ DIGNITATIS 1343

Structure of the presentation

- Problem statement
- Types of variables
- Methodology
- Results presented by the authors (optional)
- Simulation results
- Conclusion

Problem statement

Data analysis problems often involve pre-processing raw data, which is a tedious and time-demanding task due to several reasons:

1. Raw data is often unstructured and large- scale
2. It contains errors and missing values
3. Documentation may be incomplete or not available

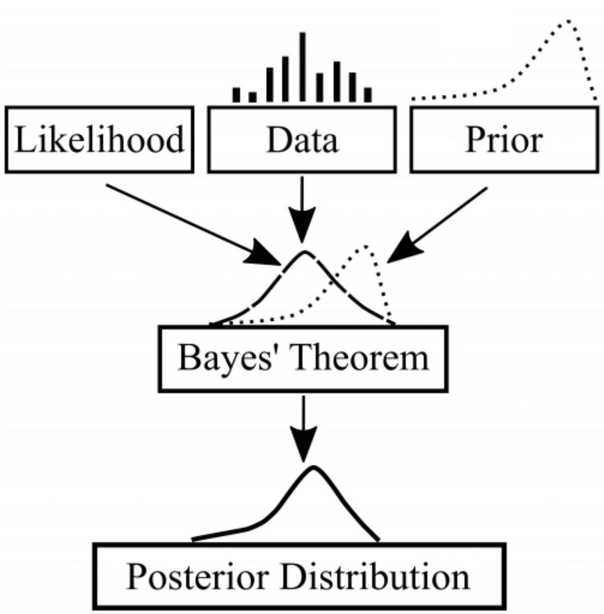
Goal: find an appropriate algorithm to **automatically recognize the datatype of the variables.**

Types of variables

Temperature (Celsius)	Temperature (bin):	Height (cm):	Car Colours:	Education Levels:	Number of vacations per year
25.6	20.0	168.5	Red	High School Diploma	3
-3.5	-10.0	182.0	Blue	Associate's Degree	1
12.0	10.0	155.2	Green	Bachelor's Degree	5
4.2	0.0	176.8	Black	Master's Degree	0
32.7	30.0	162.3	White	Doctorate Degree	2
Continuous variables			Discrete variables		
Real-valued data	Interval data	Positive real-valued data	Categorical data	Ordinal data	Count data

Methodology

Bayesian method



Likelihood functions

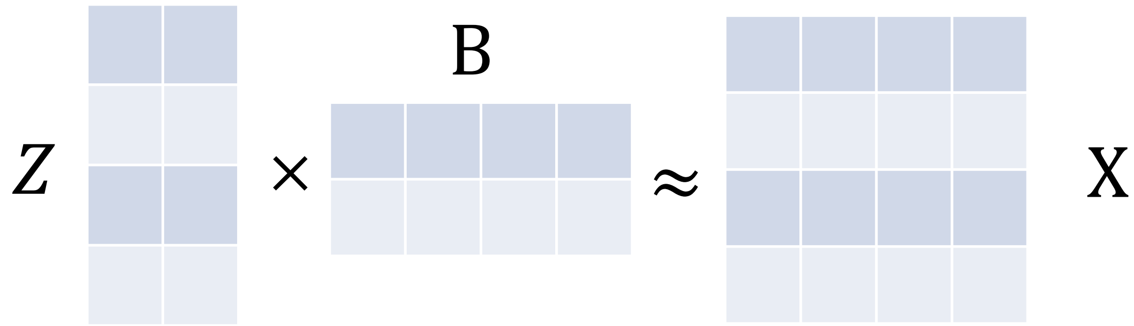
For discrete random variables:

$$\mathcal{L}(\theta \mid x) = p_\theta(x) = P_\theta(X = x)$$

For continuous random variables:

$$\mathcal{L}(\theta \mid x) = f_\theta(x)$$

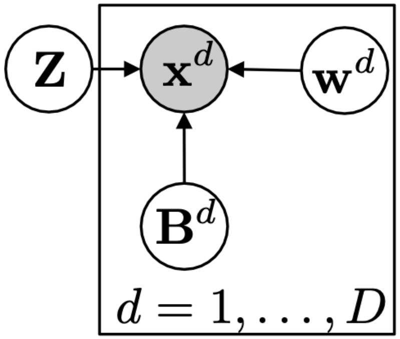
Methodology



- X: original dataset
- Z: low-rank representation of X
- B: matrix of column vectors b^d.
- b^d: weighting vector of the contribution of the latent variables to the variables of the original dataset.

Proposed model

$$p(\mathbf{x}^d | \mathbf{Z}, \{\boldsymbol{b}_\ell^d\}_{\ell \in \mathcal{L}^d}) = \sum_{\ell \in \mathcal{L}^d} w_\ell^d \, p_\ell(\mathbf{x}^d | \mathbf{Z}, \boldsymbol{b}_\ell^d)$$



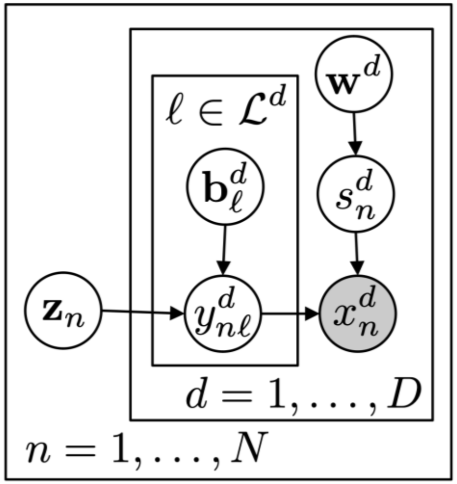
- w^d: vector of likelihood weights
- D: number of attributes of X

Goal: to find the w^d's

Possibles issues of the model:

- Different kind of likelihood function depending on the nature of the variable.
- Combination of likelihood functions with different supports.

Alternative model



- N : number of rows of X
- D : number of columns of X
- x_n^d: observation from the original data
- y_{nl}^d: pseudo-observation associated to x_n^d
- $\mathcal{L}^d$: set of possible datatypes
- s_n^d: likelihood assignments

Passing from the pseudo-observations to the original data

For every datatype is considered a transformation function over the pseudo-observations, which maps the real line into the support of the corresponding likelihood function.

For continuous variables, the transformations are:

- For real-valued data, $f_R(y) = \omega * y + \mu$
- For positive real-valued data, $f_R(y) = \ln(1 + e^{\omega * y})$
- For interval data $f_R(y) = \frac{\theta_H - \theta_L}{1 + e^{-\omega * y}} - \theta_L$

Using the likelihood functions

Likelihood functions are used in the computation of the likelihood assignments S.

$$p(s_n^d = \ell | \mathbf{w}^d, \mathbf{Z}, \{\boldsymbol{b}_\ell^d\}) = \frac{w_\ell^d \, p_\ell(x_n^d | \boldsymbol{z}_n, \boldsymbol{b}_\ell^d)}{\sum_{\ell' \in \mathcal{L}^d} w_{\ell'}^d \, p_{\ell'}(x_n^d | \boldsymbol{z}_n, \boldsymbol{b}_{\ell'}^d)}.$$

Where the likelihood functions for continuous data are defined as:

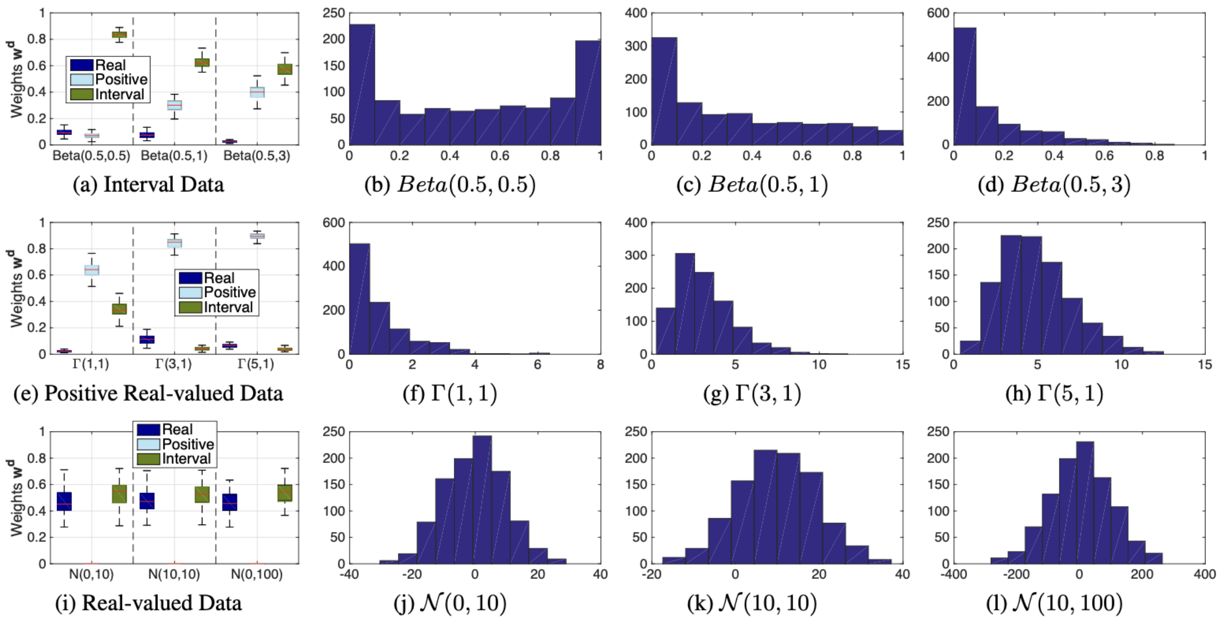
$$p_\ell(x_n^d | \mathbf{z}_n, \boldsymbol{b}_\ell^d, s_n^d = \ell) = \frac{1}{\sqrt{2\pi(\sigma_y^2 + \sigma_u^2)}} \left| \frac{d}{dx_n^d} f_\ell^{-1}(x_n^d) \right|$$
$$\times \exp\left\{ -\frac{1}{2(\sigma_y^2 + \sigma_u^2)} (f_\ell^{-1}(x_n^d) - \mathbf{z}_n \boldsymbol{b}_\ell^d)^2 \right\},$$

Pseudo-code of the algorithm

Algorithm 1: Inference Algorithm

Input: X;
Initialize: S, $\{b_l^d\}$ and $\{y_{nl}^d\}$;
for *each iteration* **do**
    Update Z given $\{b_l^d\}$ and $\{y_{nl}^d\}$;
    for $d = 1, \ldots, D$ **do**
        for $l \in \mathcal{L}^d$ **do**
            for $n = 1, \ldots, N$ **do**
                Sample $\{y_{nl}^d\}$ given x_n^d, Z, $\{b_l^d\}$, and s_n^d;
            end
            Sample $\{b_l^d\}$ given Z, and $\{y_{nl}^d\}$;
        end
        for $n = 1, , N$ **do**
            Sample s_n^d given x_n^d, Z and $\{b_l^d\}$;
        end
        Sample w^d given S;
    end
end
Result: Likelihood weights w^d

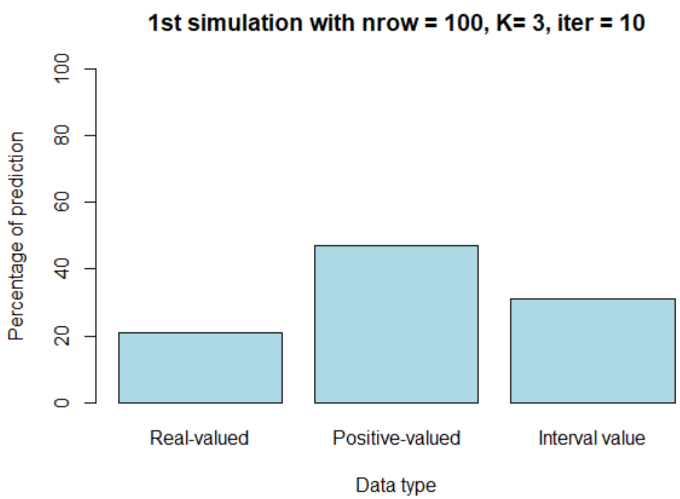
Results from the paper

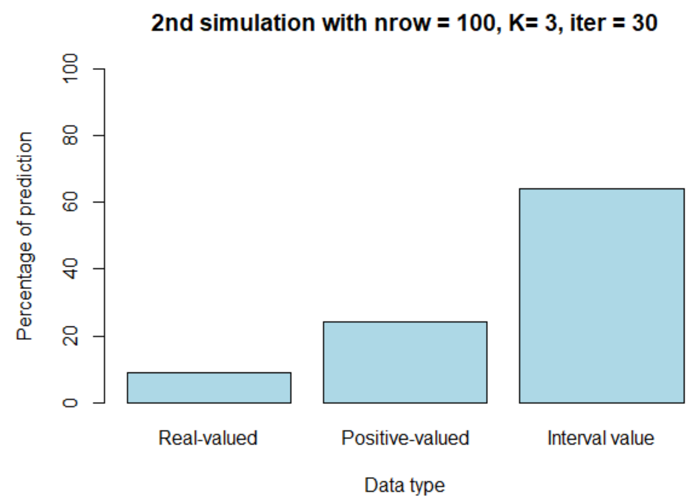


The dataset used had the following characteristics:

- 1000 observations
- Real-valued data generated with Gaussian distribution
- Positive real-valued data generated with Gamma distribution
- Interval data generated with Beta distribution

Implementation of the pseudo-code in R and some simulations





Two simulations were implemented with the following parameters:

- Number of iterations per simulation: 100.
- Each iteration used a dataset randomly generated with 100 rows and 3 columns. One column per data type.

Conclusion

- The performance of the algorithm depends on the size of the dataset.
- Computationally expensive due to the computation of the inverse of several matrices.
- The number of variables of the low rank representation and the number of iterations of the algorithm influence the result.