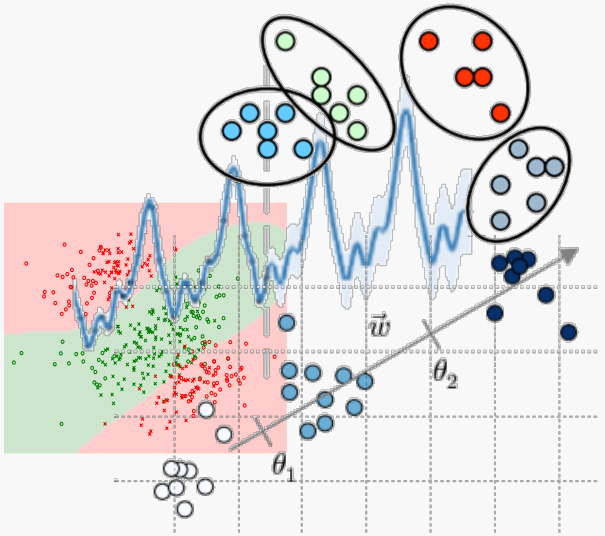
Automatic Discovery of the Statistical Types of Variables in a Dataset

Isabel Valera &
Zoubin Ghahramani



DALI 2017

Tenerife, 18th April2017

Motivation

Data from a wide range of scenarios & sources:

Medical data

Online data

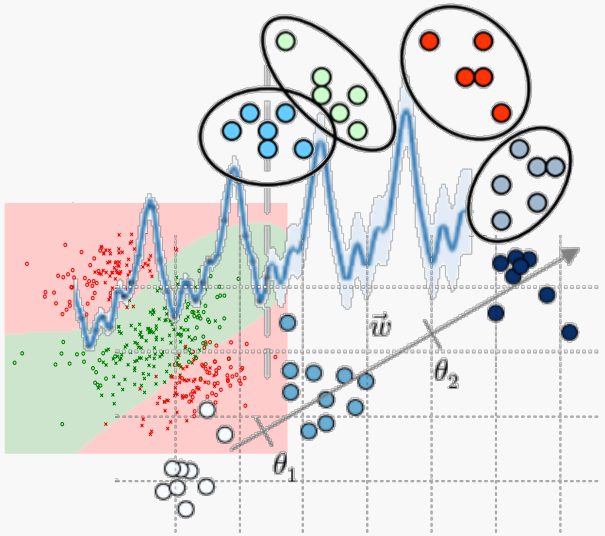
Sensor data

Motivation

Attributes

Objects

	1	2	3	4	5	6	7	8
1	1	14.2300	1.7100	2.4300	15.6000	58	2.8000	3.0600
2	1	13.2000	1.7800	2.1400	11.2000	31	2.6500	2.7600
3	1	13.1600	2.3600	2.6700	18.6000	32	2.8000	3.2400
4	1	14.3700	1.9500	2.5000	16.8000	44	3.8500	3.4900
5	1	13.2400	2.5900	2.8700	21	49	2.8000	2.6900
6	1	14.2000	1.7600	2.4500	15.2000	43	3.2700	3.3900
7	1	14.3900	1.8700	2.4500	14.6000	27	2.5000	2.5200
8	1	14.0600	2.1500	2.6100	17.6000	52	2.6000	2.5100
9	1	14.8300	1.6400	2.1700	14	28	2.8000	2.9800
10	1	13.8600	1.3500	2.2700	16	29	2.9800	3.1500
11	1	14.1000	2.1600	2.3000	18	36	2.9500	3.3200
12	1	14.1200	1.4800	2.3200	16.8000	26	2.2000	2.4300
13	1	13.7500	1.7300	2.4100	16	20	2.6000	2.7600
14	1	14.7500	1.7300	2.3900	11.4000	22	3.1000	3.6900
15	1	14.3800	1.8700	2.3800	12	33	3.3000	3.6400
16	1	13.6300	1.8100	2.7000	17.2000	43	2.8500	2.9100
17	1	14.3000	1.9200	2.7200	20	51	2.8000	3.1400
18	1	13.8300	1.5700	2.6200	20	46	2.9500	3.4000
19	1	14.1900	1.5900	2.4800	16.5000	39	3.3000	3.9300
20	1	13.6400	3.1000	2.5600	15.2000	47	2.7000	3.0300
21	1	14.0600	1.6300	2.2800	16	57	3	3.1700
22	1	12.9300	3.8000	2.6500	18.6000	33	2.4100	2.4100
23	1	13.7100	1.8600	2.3600	16.6000	32	2.6100	2.8800
24	1	12.8500	1.6000	2.5200	17.8000	26	2.4800	2.3700
25	1	13.5000	1.8100	2.6100	20	27	2.5300	2.6100



Raw data

- Unstructured
- Errors/missing
- Improper or incomplete documentation

Pre-process

Dataset

- Structured
- Objects/ attributes

Labeling

Predictive Analytics

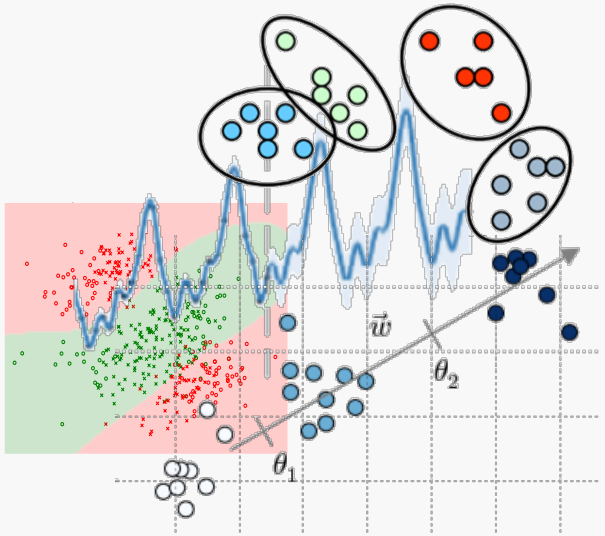
- Task: regression, classification, clustering, etc.
- Likelihood model

Motivation

Attributes

Objects

	1	2	3	4	5	6	7	8
1	1	14.2300	1.7100	2.4300	15.6000	58	2.8000	3.0600
2	1	13.2000	1.7800	2.1400	11.2000	31	2.6500	2.7600
3	1	13.1600	2.3600	2.6700	18.6000	32	2.8000	3.2400
4	1	14.3700	1.9500	2.5000	16.8000	44	3.8500	3.4900
5	1	13.2400	2.5900	2.8700	21	49	2.8000	2.6900
6	1	14.2000	1.7600	2.4500	15.2000	43	3.2700	3.3900
7	1	14.3900	1.8700	2.4500	14.6000	27	2.5000	2.5200
8	1	14.0600	2.1500	2.6100	17.6000	52	2.6000	2.5100
9	1	14.8300	1.6400	2.1700	14	28	2.8000	2.9800
10	1	13.8600	1.3500	2.2700	16	29	2.9800	3.1500
11	1	14.1000	2.1600	2.3000	18	36	2.9500	3.3200
12	1	14.1200	1.4800	2.3200	16.8000	26	2.2000	2.4300
13	1	13.7500	1.7300	2.4100	16	20	2.6000	2.7600
14	1	14.7500	1.7300	2.3900	11.4000	22	3.1000	3.6900
15	1	14.3800	1.8700	2.3800	12	33	3.3000	3.6400
16	1	13.6300	1.8100	2.7000	17.2000	43	2.8500	2.9100
17	1	14.3000	1.9200	2.7200	20	51	2.8000	3.1400
18	1	13.8300	1.5700	2.6200	20	46	2.9500	3.4000
19	1	14.1900	1.5900	2.4800	16.5000	39	3.3000	3.9300
20	1	13.6400	3.1000	2.5600	15.2000	47	2.7000	3.0300
21	1	14.0600	1.6300	2.2800	16	57	3	3.1700
22	1	12.9300	3.8000	2.6500	18.6000	33	2.4100	2.4100
23	1	13.7100	1.8600	2.3600	16.6000	32	2.6100	2.8800
24	1	12.8500	1.6000	2.5200	17.8000	26	2.4800	2.3700
25	1	13.5000	1.8100	2.6100	20	27	2.5300	2.6100



Raw data → **Pre-process** → **Dataset** → **Labeling** → **Predictive Analytics**

Tools for **automatic**

i) data cleaning ✔

ii) data wrangling ✔

iii) data integration ✔

Data types (and often likelihood models) assumed as known.

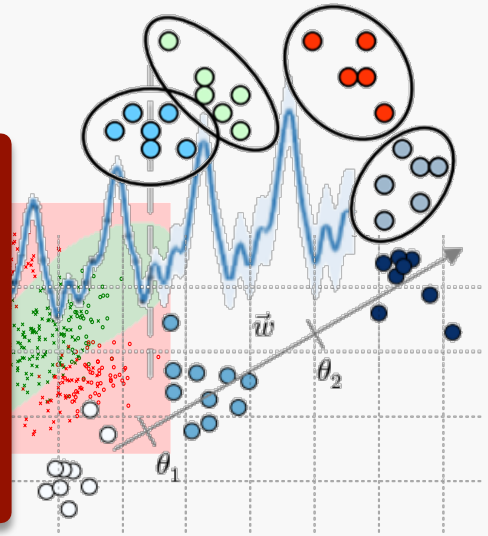
Lack of tools for automatic data labeling!

Motivation

Attributes

	1	2	3	4	5	6	7	8
1	1	14.2300	1.7100	2.4300	15.6000	58	2.8000	3.0600
2	1	13.2000	1.7800	2.1400	11.2000	31	2.6500	2.7600
3	1	13.1600	2.3600	2.6700	18.6000	32	2.8000	3.2400

Bayesian Method to Discover the Statistical Types of Data



Raw data → Pre-process → **Dataset** → Labeling → **Predictive Analytics**

Tools for **automatic**
- i) data cleaning ✔
- ii) data wrangling ✔
- iii) data integration ✔

Data types (and often likelihood models) assumed as known.

Lack of tools for automatic data labeling!

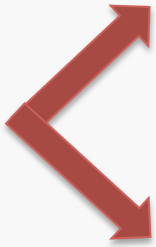
Problem Statement

Attributes

Objects

	1	2	3	4	5	6	7	8
1	1	14.2300	1.7100	2.4300	15.6000	58	2.8000	3.0600
2	1	13.2000	1.7800	2.1400	11.2000	31	2.6500	2.7600
3	1	13.1600	2.3600	2.6700	18.6000	32	2.8000	3.2400
4	1	14.3700	1.9500	2.5000	16.8000	44	3.8500	3.4900
5	1	13.2400	2.5900	2.8700	21	49	2.8000	2.6900
6	1	14.2000	1.7600	2.4500	15.2000	43	3.2700	3.3900
7	1	14.3900	1.8700	2.4500	14.6000	27	2.5000	2.5200
8	1	14.0600	2.1500	2.6100	17.6000	52	2.6000	2.5100
9	1	14.8300	1.6400	2.1700	14	28	2.8000	2.9800
10	1	13.8600	1.3500	2.2700	16	29	2.9800	3.1500
11	1	14.1000	2.1600	2.3000	18	36	2.9500	3.3200
12	1	14.1200	1.4800	2.3200	16.8000	26	2.2000	2.4300
13	1	13.7500	1.7300	2.4100	16	20	2.6000	2.7600
14	1	14.7500	1.7300	2.3900	11.4000	22	3.1000	3.6900
15	1	14.3800	1.8700	2.3800	12	33	3.3000	3.6400
16	1	13.6300	1.8100	2.7000	17.2000	43	2.8500	2.9100
17	1	14.3000	1.9200	2.7200	20	51	2.8000	3.1400
18	1	13.8300	1.5700	2.6200	20	46	2.9500	3.4000
19	1	14.1900	1.5900	2.4800	16.5000	39	3.3000	3.9300
20	1	13.6400	3.1000	2.5600	15.2000	47	2.7000	3.0300
21	1	14.0600	1.6300	2.2800	16	57	3	3.1700
22	1	12.9300	3.8000	2.6500	18.6000	33	2.4100	2.4100
23	1	13.7100	1.8600	2.3600	16.6000	32	2.6100	2.8800
24	1	12.8500	1.6000	2.5200	17.8000	26	2.4800	2.3700
25	1	13.5000	1.8100	2.6100	20	27	2.5300	2.6100

Continuous



Discrete

Problem Statement

Attributes

Objects

	1	2	3	4	5	6	7	8
1	1	14.2300	1.7100	2.4300	15.6000	58	2.8000	3.0600
2	1	13.2000	1.7800	2.1400	11.2000	31	2.6500	2.7600
3	1	13.1600	2.3600	2.6700	18.6000	32	2.8000	3.2400
4	1	14.3700	1.9500	2.5000	16.8000	44	3.8500	3.4900
5	1	13.2400	2.5900	2.8700	21	49	2.8000	2.6900
6	1	14.2000	1.7600	2.4500	15.2000	43	3.2700	3.3900
7	1	14.3900	1.8700	2.4500	14.6000	27	2.5000	2.5200
8	1	14.0600	2.1500	2.6100	17.6000	52	2.6000	2.5100
9	1	14.8300	1.6400	2.1700	14	28	2.8000	2.9800
10	1	13.8600	1.3500	2.2700	16	29	2.9800	3.1500
11	1	14.1000	2.1600	2.3000	18	36	2.9500	3.3200
12	1	14.1200	1.4800	2.3200	16.8000	26	2.2000	2.4300
13	1	13.7500	1.7300	2.4100	16	20	2.6000	2.7600
14	1	14.7500	1.7300	2.3900	11.4000	22	3.1000	3.6900
15	1	14.3800	1.8700	2.3800	12	33	3.3000	3.6400
16	1	13.6300	1.8100	2.7000	17.2000	43	2.8500	2.9100
17	1	14.3000	1.9200	2.7200	20	51	2.8000	3.1400
18	1	13.8300	1.5700	2.6200	20	46	2.9500	3.4000
19	1	14.1900	1.5900	2.4800	16.5000	39	3.3000	3.9300
20	1	13.6400	3.1000	2.5600	15.2000	47	2.7000	3.0300
21	1	14.0600	1.6300	2.2800	16	57	3	3.1700
22	1	12.9300	3.8000	2.6500	18.6000	33	2.4100	2.4100
23	1	13.7100	1.8600	2.3600	16.6000	32	2.6100	2.8800
24	1	12.8500	1.6000	2.5200	17.8000	26	2.4800	2.3700
25	1	13.5000	1.8100	2.6100	20	27	2.5300	2.6100

Continuous

- Real-valued
- Positive real-valued
- Interval

Discrete

- Categorical
- Ordinal
- Count

Problem Statement

Attributes

Objects

	1	2	3	4	5	6	7	8
1	1	14.2300	1.7100	2.4300	15.6000	58	2.8000	3.0600
2	1	13.2000	1.7800	2.1400	11.2000	31	2.6500	2.7600
3	1	13.1600	2.3600	2.6700	18.6000	32	2.8000	3.2400
4	1	14.3700	1.9500	2.5000	16.8000	44	3.8500	3.4900
5	1	13.2400	2.5900	2.8700	21	49	2.8000	2.6900
6	1	14.2000	1.7600	2.4500	15.2000	43	3.2700	3.3900
7	1	14.3900	1.8700	2.4500	14.6000	27	2.5000	2.5200
8	1	14.0600	2.1500	2.6100	17.6000	52	2.6000	2.5100
9	1	14.8300	1.6400	2.1700	14	28	2.8000	2.9800
10	1	13.8600	1.3500	2.2700	16	29	2.9800	3.1500
11	1	14.1000	2.1600	2.3000	18	36	2.9500	3.3200
12	1	14.1200	1.4800	2.3200	16.8000	26	2.2000	2.4300
13	1	13.7500	1.7300	2.4100	16	20	2.6000	2.7600
14	1	14.7500	1.7300	2.3900	11.4000	22	3.1000	3.6900
15	1	14.3800	1.8700	2.3800	12	33	3.3000	3.6400
16	1	13.6300	1.8100	2.7000	17.2000	43	2.8500	2.9100
17	1	14.3000	1.9200	2.7200	20	51	2.8000	3.1400
18	1	13.8300	1.5700	2.6200	20	46	2.9500	3.4000
19	1	14.1900	1.5900	2.4800	16.5000	39	3.3000	3.9300
20	1	13.6400	3.1000	2.5600	15.2000	47	2.7000	3.0300
21	1	14.0600	1.6300	2.2800	16	57	3	3.1700
22	1	12.9300	3.8000	2.6500	18.6000	33	2.4100	2.4100
23	1	13.7100	1.8600	2.3600	16.6000	32	2.6100	2.8800
24	1	12.8500	1.6000	2.5200	17.8000	26	2.4800	2.3700
25	1	13.5000	1.8100	2.6100	20	27	2.5300	2.6100

Continuous

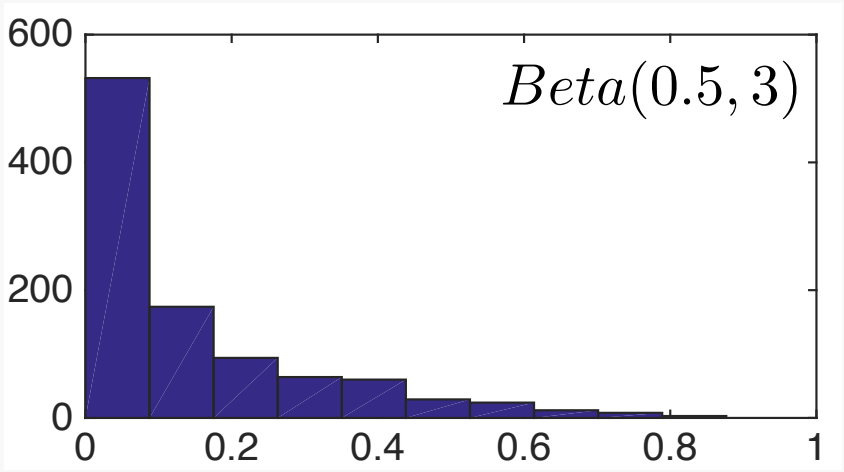
- Real-valued
- Positive real-valued
- Interval

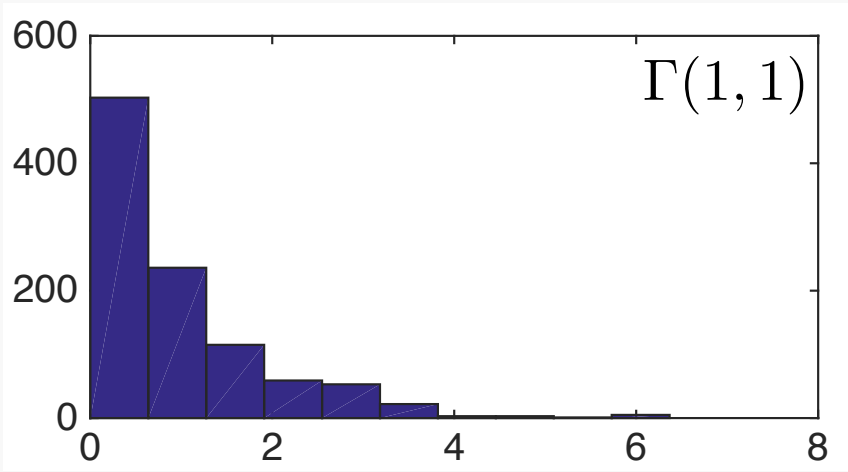
Discrete

- Categorical
- Ordinal
- Count

Main challenges:

1. Finite observed sample size





Problem Statement

Attributes

Objects

	1	2	3	4	5	6	7	8
1	1	14.2300	1.7100	2.4300	15.6000	58	2.8000	3.0600
2	1	13.2000	1.7800	2.1400	11.2000	31	2.6500	2.7600
3	1	13.1600	2.3600	2.6700	18.6000	32	2.8000	3.2400
4	1	14.3700	1.9500	2.5000	16.8000	44	3.8500	3.4900
5	1	13.2400	2.5900	2.8700	21	49	2.8000	2.6900
6	1	14.2000	1.7600	2.4500	15.2000	43	3.2700	3.3900
7	1	14.3900	1.8700	2.4500	14.6000	27	2.5000	2.5200
8	1	14.0600	2.1500	2.6100	17.6000	52	2.6000	2.5100
9	1	14.8300	1.6400	2.1700	14	28	2.8000	2.9800
10	1	13.8600	1.3500	2.2700	16	29	2.9800	3.1500
11	1	14.1000	2.1600	2.3000	18	36	2.9500	3.3200
12	1	14.1200	1.4800	2.3200	16.8000	26	2.2000	2.4300
13	1	13.7500	1.7300	2.4100	16	20	2.6000	2.7600
14	1	14.7500	1.7300	2.3900	11.4000	22	3.1000	3.6900
15	1	14.3800	1.8700	2.3800	12	33	3.3000	3.6400
16	1	13.6300	1.8100	2.7000	17.2000	43	2.8500	2.9100
17	1	14.3000	1.9200	2.7200	20	51	2.8000	3.1400
18	1	13.8300	1.5700	2.6200	20	46	2.9500	3.4000
19	1	14.1900	1.5900	2.4800	16.5000	39	3.3000	3.9300
20	1	13.6400	3.1000	2.5600	15.2000	47	2.7000	3.0300
21	1	14.0600	1.6300	2.2800	16	57	3	3.1700
22	1	12.9300	3.8000	2.6500	18.6000	33	2.4100	2.4100
23	1	13.7100	1.8600	2.3600	16.6000	32	2.6100	2.8800
24	1	12.8500	1.6000	2.5200	17.8000	26	2.4800	2.3700
25	1	13.5000	1.8100	2.6100	20	27	2.5300	2.6100

Continuous

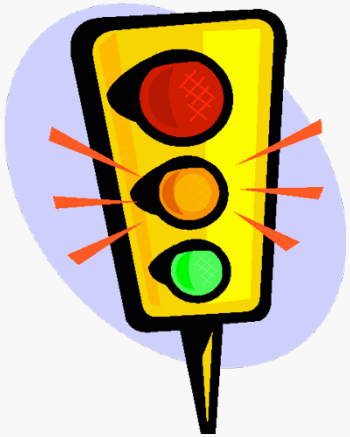
- Real-valued
- Positive real-valued
- Interval

Discrete

- Categorical
- Ordinal
- Count

Main challenges:

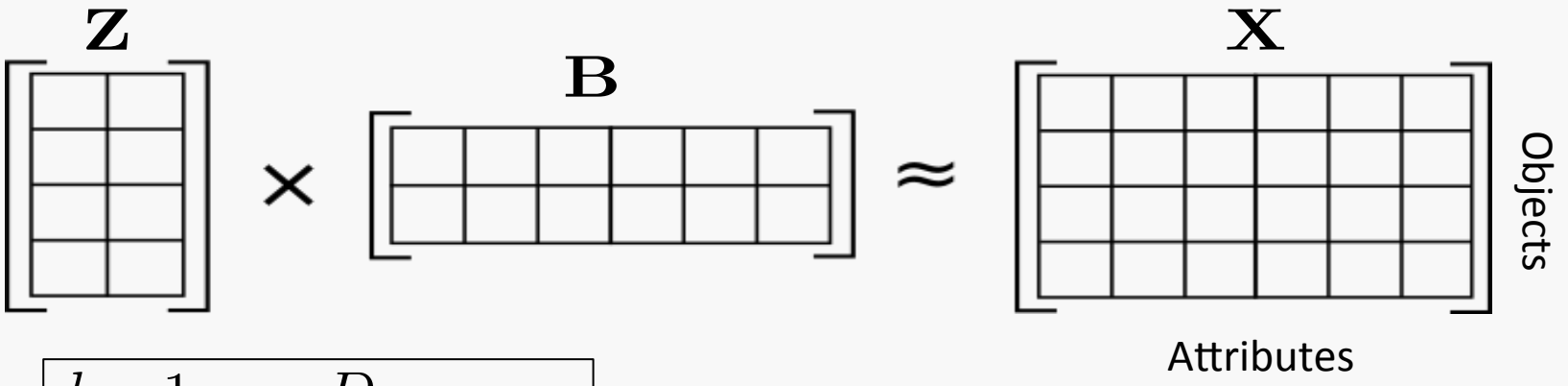
2. Order makes sense only given a context

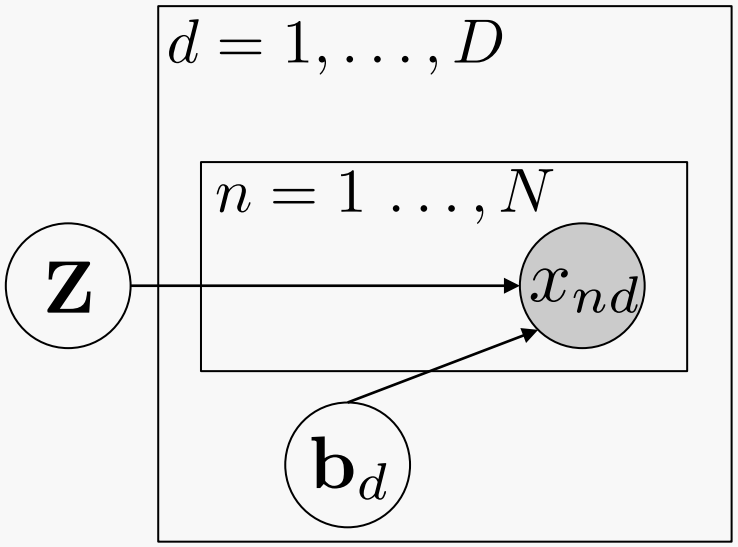


Automatic Discovery of Data types

Key ideas:

i) Low-rank representation





$$p(\mathbf{X}|\mathbf{Z}, \mathbf{B}) = \prod_{d=1}^{D} p(\mathbf{x}_d|\mathbf{Z}, \mathbf{b}_d)$$

Automatic Discovery of Data types

Key ideas:

i) Low-rank representation

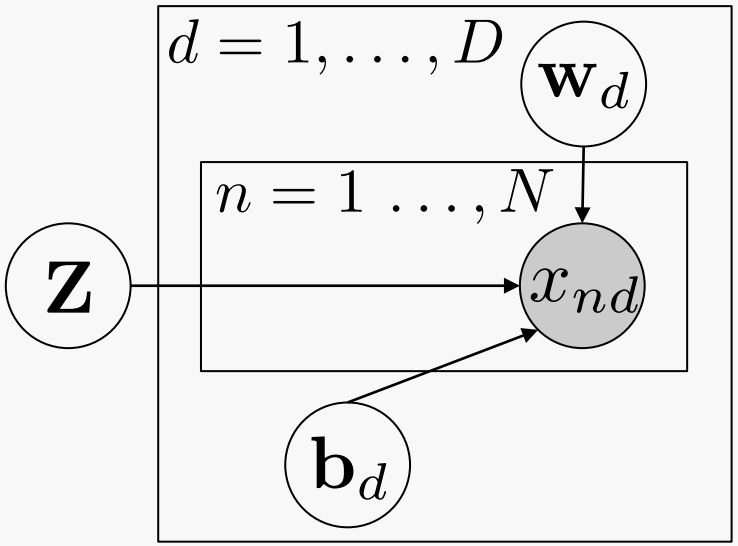
ii) Likelihood as a mixture model

Weight of the ℓ-th likelihood model:

$$\sum_{\ell=1}^{L} w_d(\ell) = 1$$

$$p(\mathbf{X}|\mathbf{Z},\mathbf{B}) = \prod_{d=1}^{D} \sum_{\ell=1}^{L} w_d(\ell) p_\ell(\mathbf{x}_d|\mathbf{Z},\mathbf{b}_d)$$

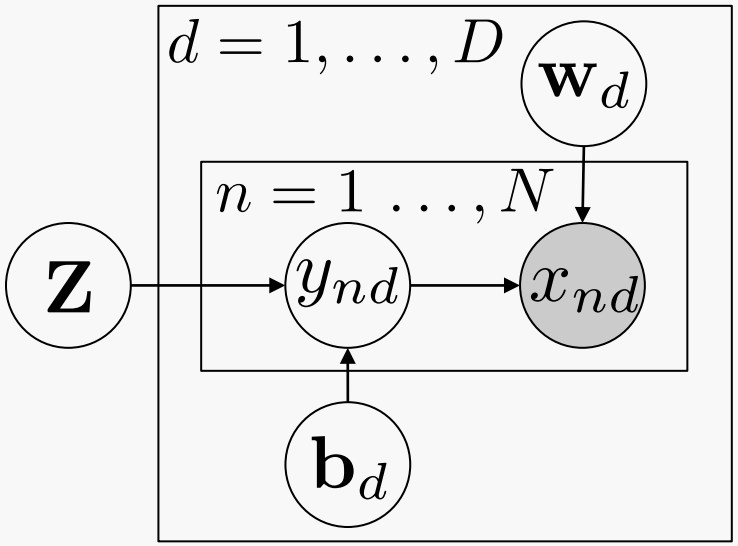
ℓ-th normalized likelihood model



Automatic Discovery of Data types

Key ideas:

i) Low-rank representation
ii) Likelihood as a mixture model
iii) Auxiliary real-valued pseudo-observation



Support of ℓ-th likelihood function

$$f_\ell : \Re \rightarrow \Omega_\ell$$

$$y \rightarrow x$$

Low-rank representation

$$y_{nd} \sim \mathcal{N}(\mathbf{z}_n \mathbf{b}_d, \sigma_y^2)$$

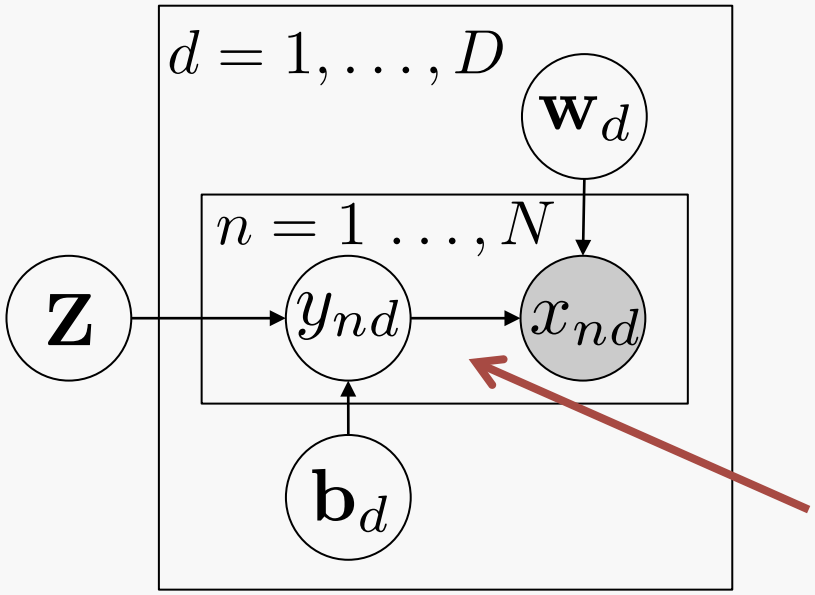
$$x_{nd} = f_\ell(y_{nd} + u_{nd})$$

Gaussian noise

Automatic Discovery of Data types

Key ideas:

i) Low-rank representation
ii) Likelihood as a mixture model
iii) Auxiliary real pseudo-observation



- MCMC inference
- Linear complexity
- Ideally $\mathbf{w}_d$ sparse

$$f_\ell : \Re \rightarrow \Omega_\ell$$

$$y \rightarrow x$$

Statistical Types of Variables

Continuous Variables:

i) Real data: $x_{nd} \in \Re$

$$x_{nd} = f_{\Re}(y_{nd} + u_{nd}) = w(y_{nd} + u_{nd}) + \mu$$

ii) Positive Real data: $x_{nd} \in \Re^{+}$

$$x_{nd} = f_{\Re^{+}}(y_{nd} + u_{nd}) = \log(1 + \exp(w(y_{nd} + u_{nd})))$$

iii) Interval data: $x_{nd} \in (\theta_L, \theta_H)$

$$x_{nd} = f_{Int}(y_{nd} + u_{nd}) = \frac{\theta_H - \theta_L}{1 + \exp(-w(y_{nd} + u_{nd}))} + \theta_L$$

Statistical Types of Variables

Discrete Variables:

i) Categorical data: $x_{nd} \in \{\text{“}blue\text{”}, \text{“}red\text{”}, \text{“}black\text{”}\}$

$$x_{nd} = f_{cat}(y_{nd}) = \arg \max_{r \in \{1,\ldots,R_d\}} y_{nd}(r),$$

ii) Ordinal data: $x_{nd} \in \{\text{“}never\text{”}, \text{“}often\text{”}, \text{“}always\text{”}\}$

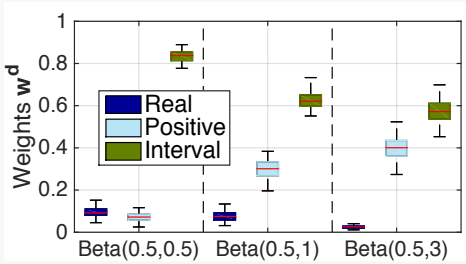
$$x_{nd} = f_{ord}(y_{nd}) = \begin{cases} 1 & \text{if } y_{nd} \leq \theta_1 \\ 2 & \text{if } \theta_1 < y_{nd} \leq \theta_2 \\ \vdots & \\ R & \text{if } \theta_{R-1} < y_{nd} \end{cases}$$

iii) Count data: $x_{nd} \in \{0, 1, \ldots, \infty\}$

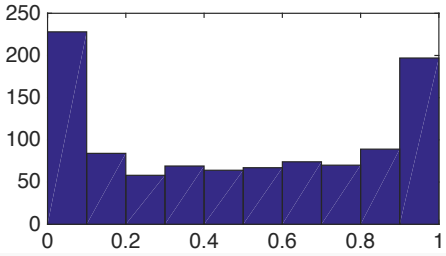
$$x_{nd} = f_{count}(y_{nd}) = \lfloor g(y_n^d) \rfloor, \text{ where } g : \Re \rightarrow \Re^+$$

Results on Continuous Variables

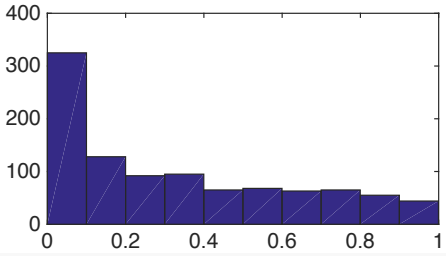
Synthetic data:



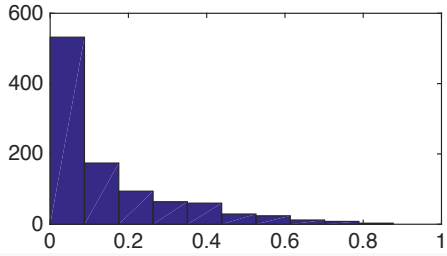
(a) Interval Data



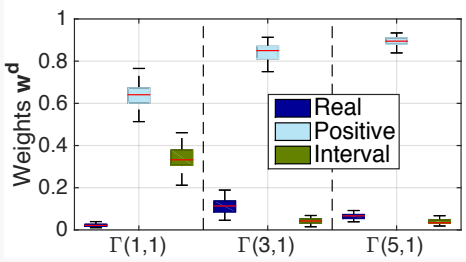
(b) $Beta(0.5, 0.5)$



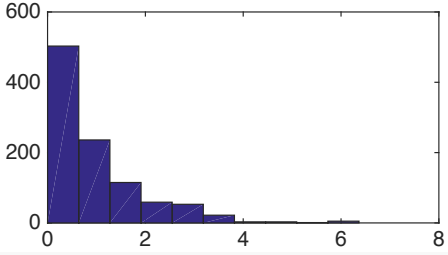
(c) $Beta(0.5, 1)$



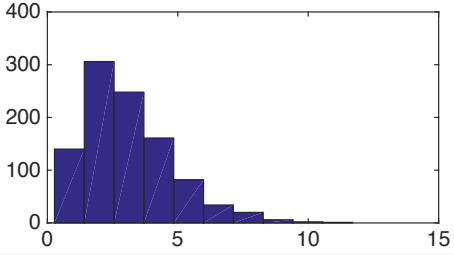
(d) $Beta(0.5, 3)$



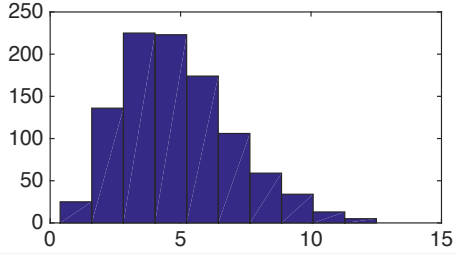
(e) Positive Real-valued Data



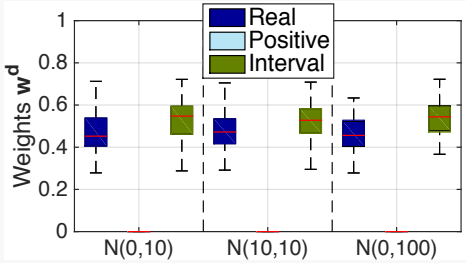
(f) $\Gamma(1, 1)$



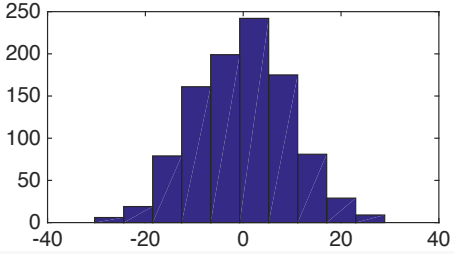
(g) $\Gamma(3, 1)$



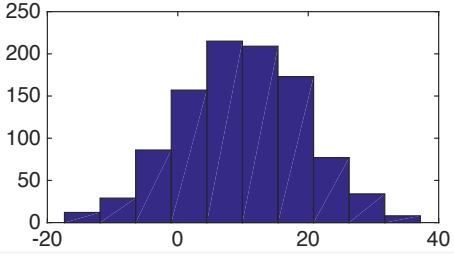
(h) $\Gamma(5, 1)$



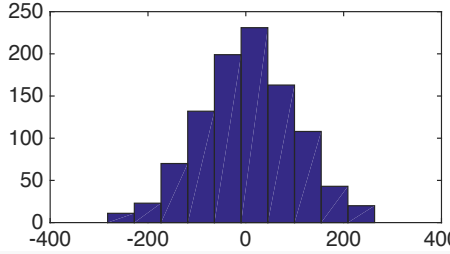
(i) Real-valued Data



(j) $\mathcal{N}(0, 10)$



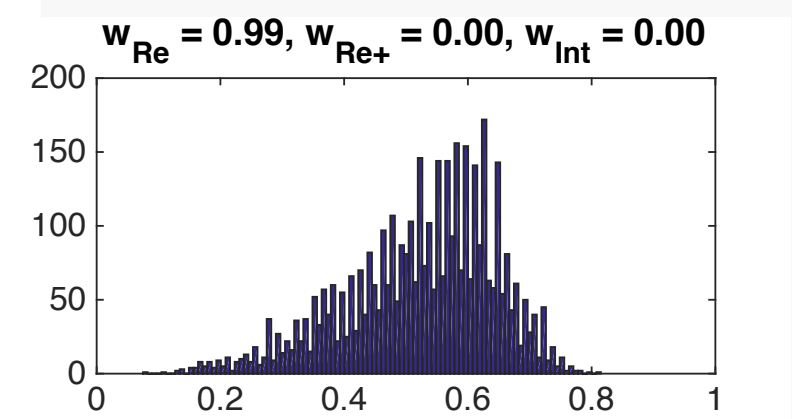
(k) $\mathcal{N}(10, 10)$



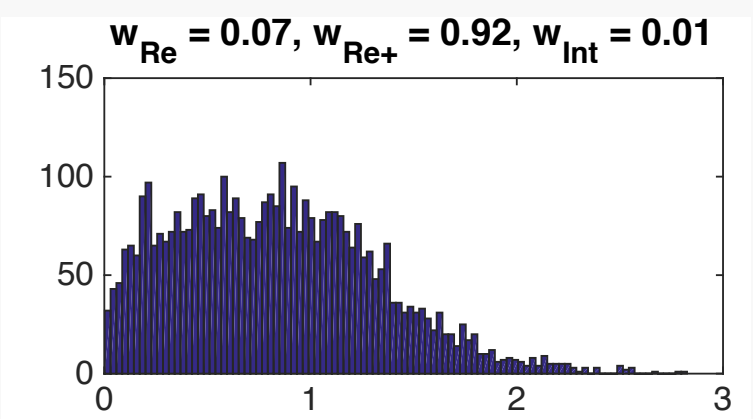
(l) $\mathcal{N}(10, 100)$

Results on Continuous Variables

Real data (Abalone dataset):



(a) Length (mm)

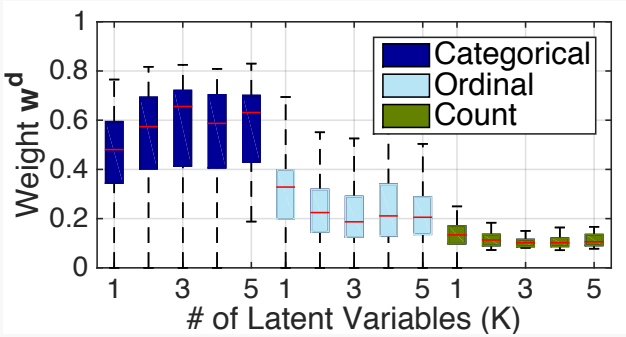


(b) Weight (grams)

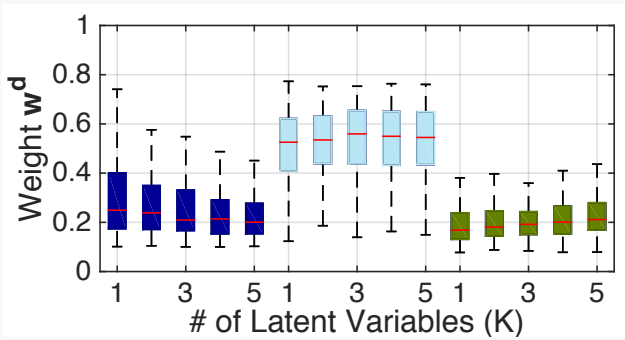
The structure in the data helps to find the most likely solution for the data

Results on Discrete Variables

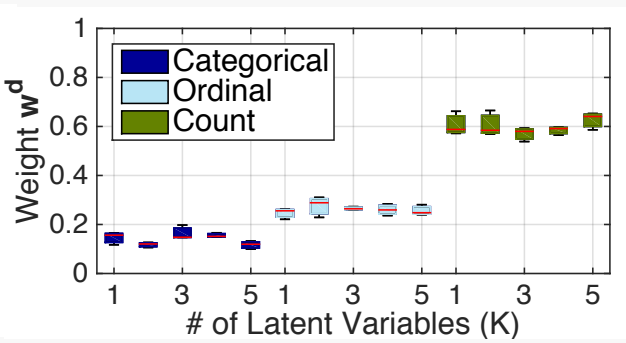
Synthetic Data:



(a) Categorical Data



(b) Ordinal Data

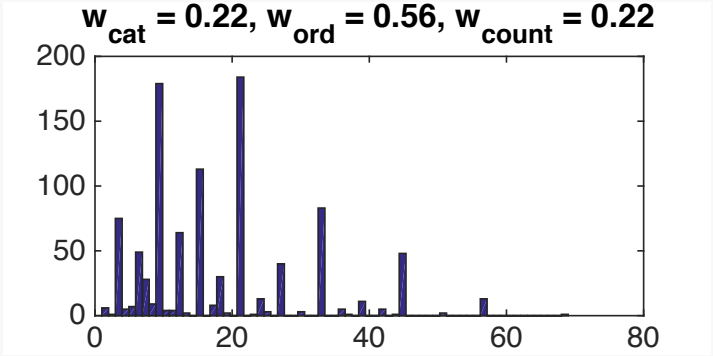


(c) Count Data

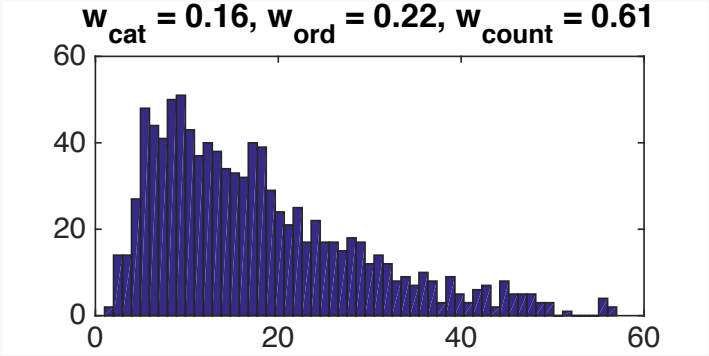
Discovery of an order in the data regardless the model complexity

Results on Discrete Variables

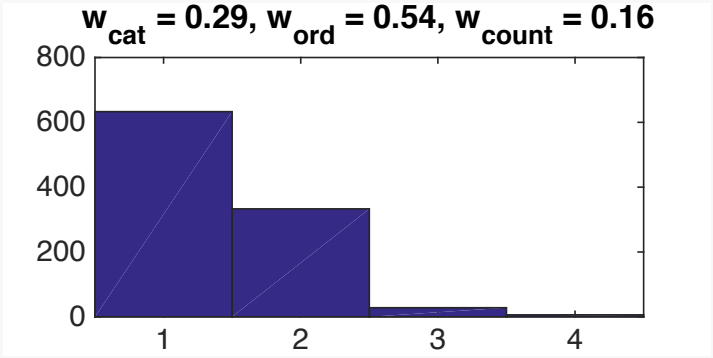
Real Data (German dataset):



(a) Duration (months)



(b) Age



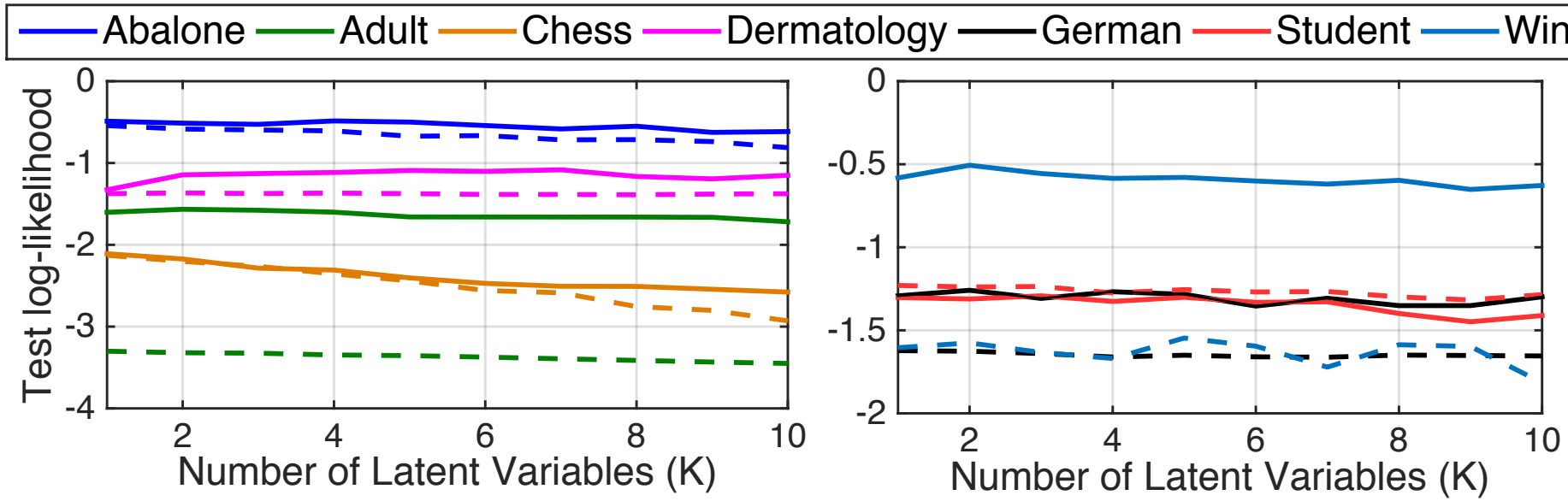
(c) #of Credits

(d) Job

Apparent failures are in fact sensible when data histograms are carefully examined

Why do we need it?

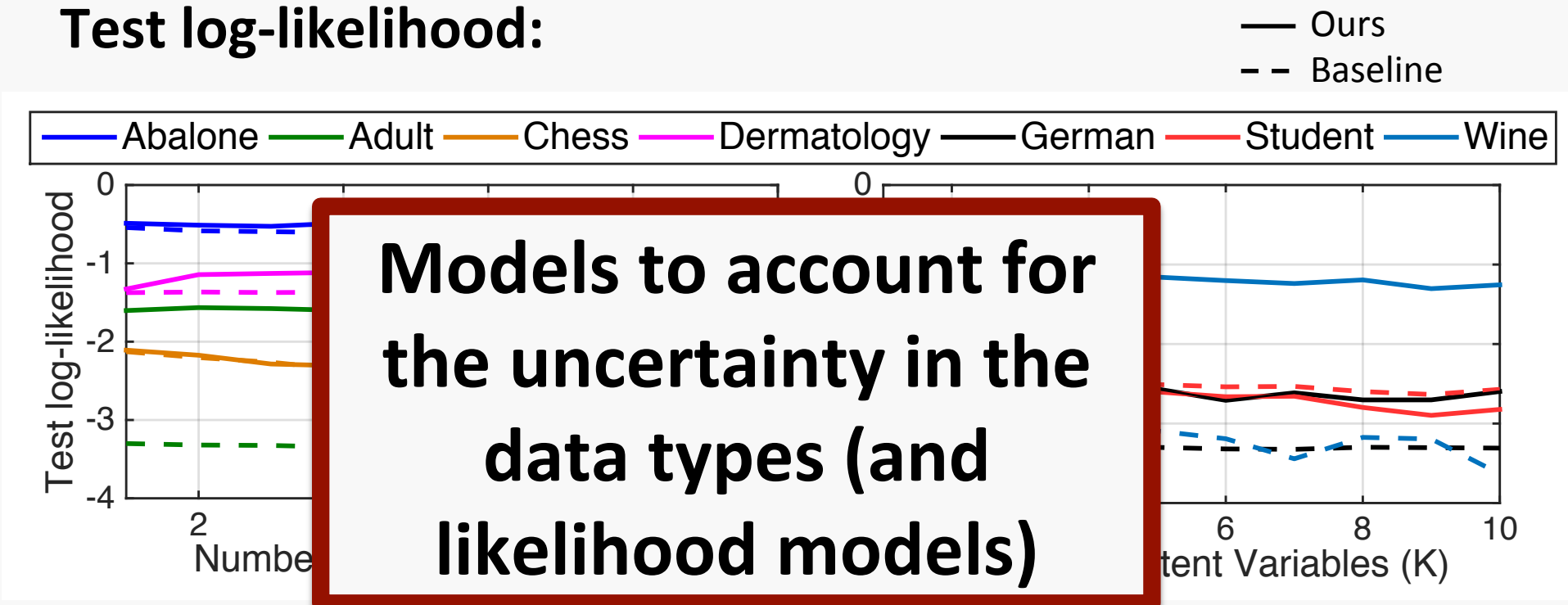
Test log-likelihood:



Taking into account the uncertainty in the statistical types of the variables, leads to a better fitting of the data

Why do we need it?

Test log-likelihood:

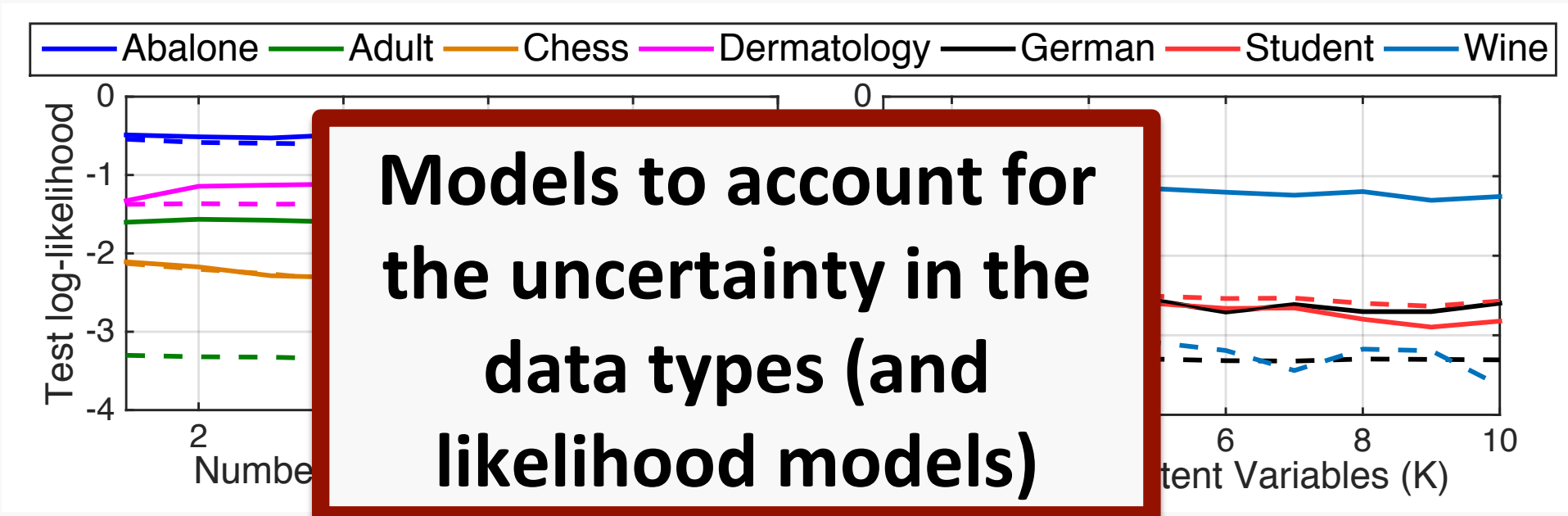


Models to account for the uncertainty in the data types (and likelihood models)

Taking into account the uncertainty in the statistical types of the variables, leads to a better fitting of the data

Why do we need it?

Test log-likelihood:



Models to account for the uncertainty in the data types (and likelihood models)

Taking into account the uncertainty in the statistical types of the variables, leads to a better fitting of the data

Thanks!