General Latent Feature Modeling for Data Exploration Tasks

Isabel Valera

University of Cambridge

Joint work with M. F. Pradier & Z. Ghaharamani

Workshop on Human Interpretability in Machine Learning. Sydney, 10th August 2017

Data exploration

???

- Before being able to solve, or even define, a predictive task we need to *explore* the data

- Helps checking assumptions required for model fitting and hypothesis testing

Data exploration

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

Standard techniques:

Visualization

+

Unsupervised Learning

- Dimensionality reductions: PCA, clustering, etc.
- Visualization of the lower-dimensional representation

Data exploration

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
- Interval

Discrete

- Categorical
- Ordinal
- Count

Challenge **– *Heterogeneous Attributes***

Example – User survey contains information on gender, age, education level, income…

Data exploration

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

Our approach:

Visualization

+

General Latent Feature Modeling

Challenge **– *Heterogeneous Attributes***

Example – User survey contains information on gender, age, education level, income…

General latent feature modeling

Suitable for data exploratory analysis?

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

- Latent structure in the data

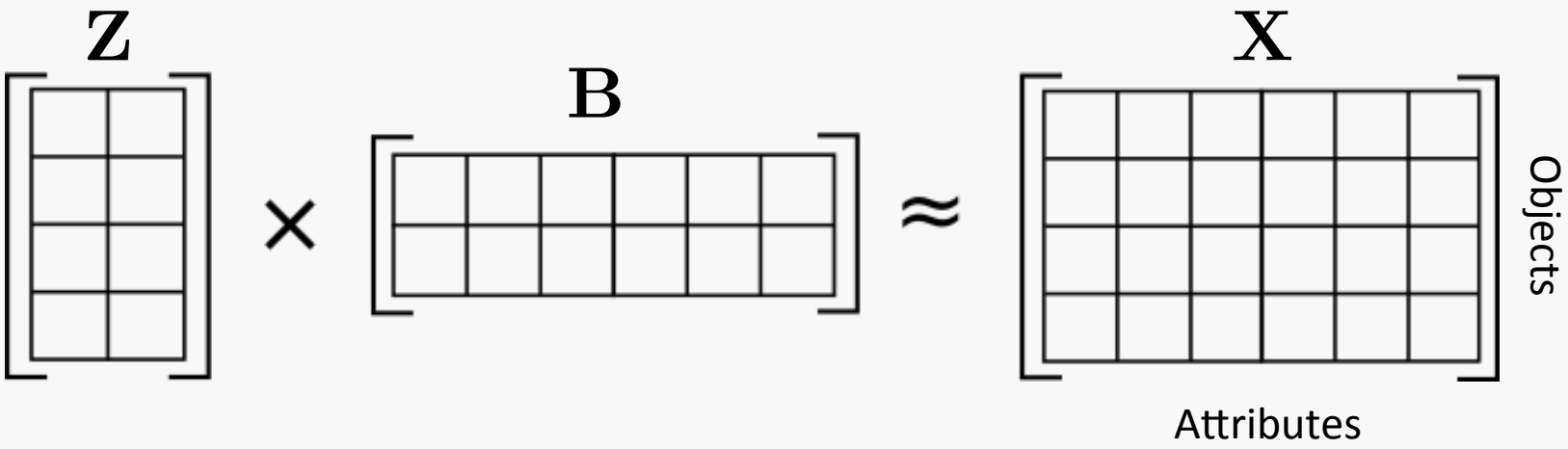
- Dependencies among objects and attributes

- Compact in a few features the immense redundant information

General latent feature modeling

How does it work?

Latent structure – Low-rank representation

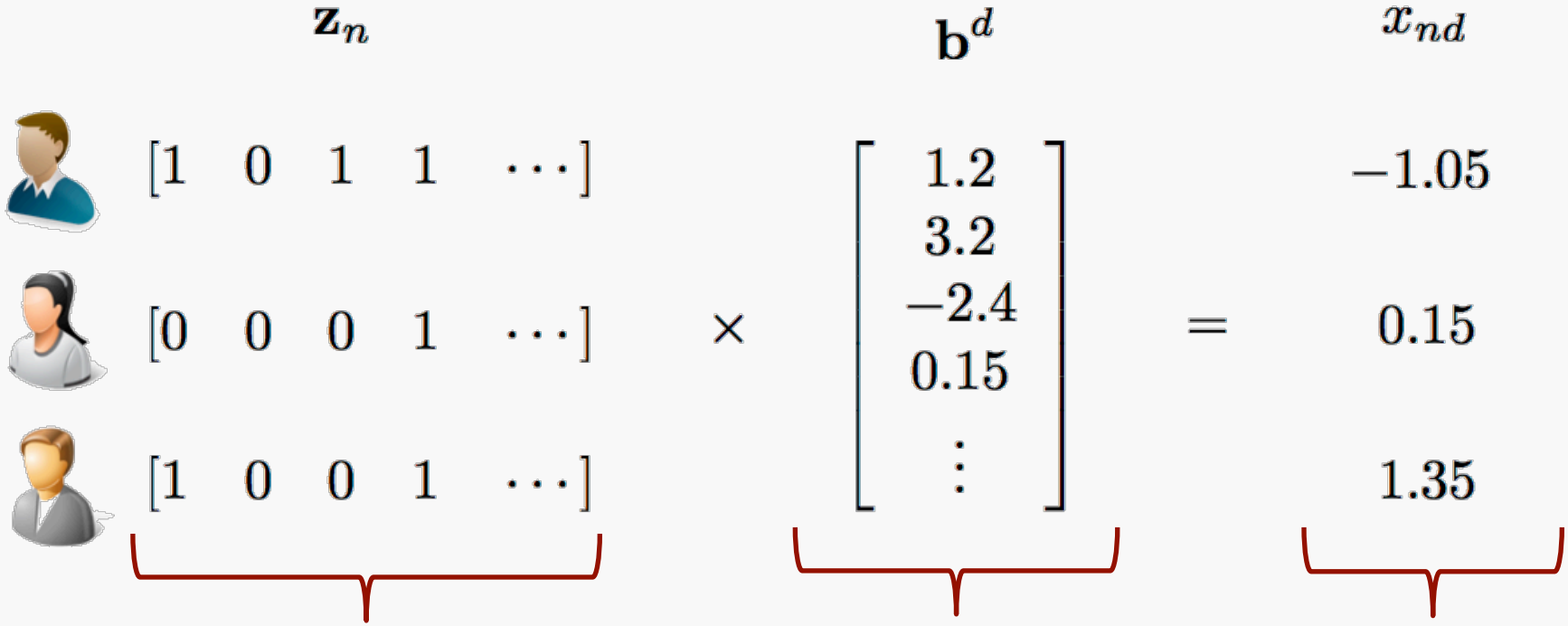


$$p(\mathbf{X}|\mathbf{Z},\mathbf{B}) = \prod_{d=1}^{D} p(\mathbf{x}_d|\mathbf{Z},\mathbf{b}_d)$$

General latent feature modeling

How does it work?

Interpretability – Objects as binary latent features



General latent feature modeling

How does it work?

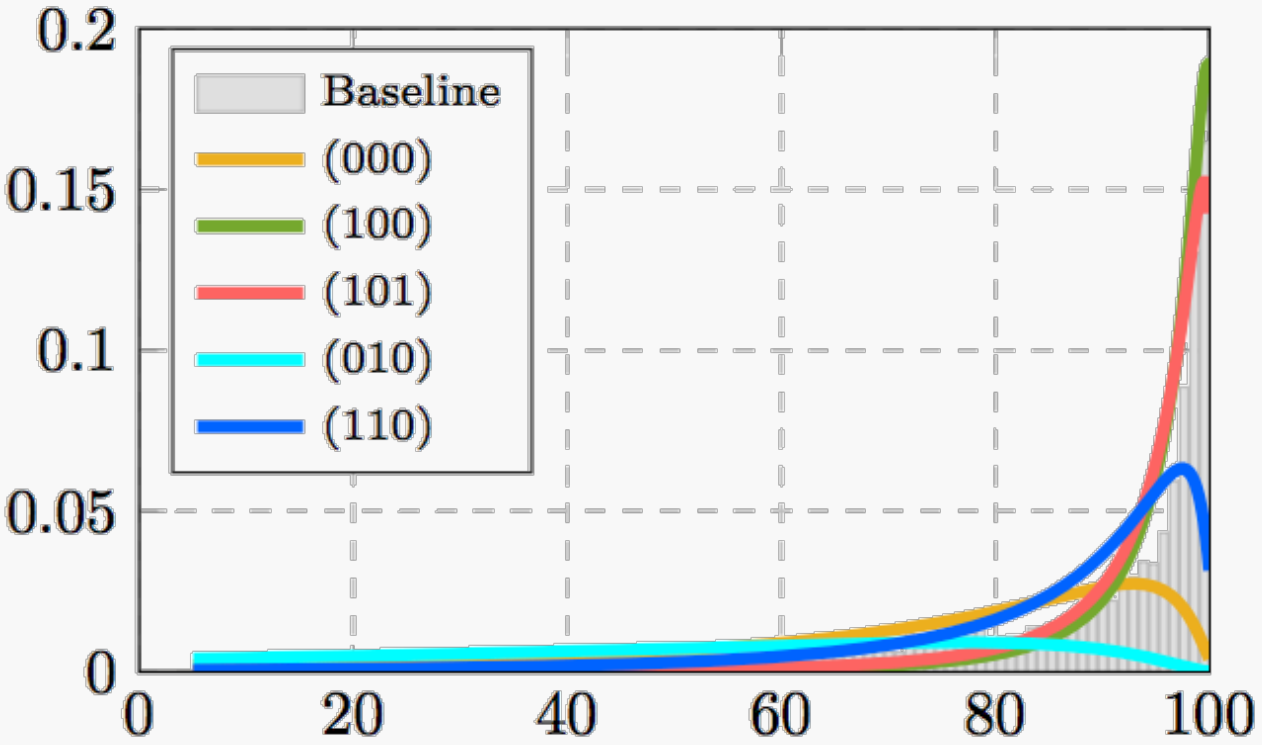
Heterogeneous data – Data transformation

$$\underbrace{\begin{matrix} \mathbf{z}_n \\ [1 \;\; 0 \;\; 1 \;\; 0 \;\; \cdots] \\ [0 \;\; 0 \;\; 0 \;\; 1 \;\; \cdots] \\ [1 \;\; 0 \;\; 0 \;\; 1 \;\; \cdots] \end{matrix} \times \begin{matrix} \mathbf{B}^d \\ \begin{bmatrix} 1.2 \\ 3.2 \\ -2.4 \\ 0.15 \\ \vdots \end{bmatrix} \end{matrix}}_{\text{Low-rank representation}} \approx \underbrace{\begin{matrix} y_n^d \\ -1.5 \\ 0.15 \\ 1.35 \end{matrix} \xrightarrow{f_d(\cdot)} \begin{matrix} x_n^d \\ 1 \\ 2 \\ 2 \end{matrix} \begin{matrix} \\ (\text{“}low\text{”}) \\ (\text{“}medium\text{”}) \\ (\text{“}medium\text{”}) \end{matrix}}_{\text{Data transformation}}$$

General latent feature modeling

How does it work?

Visualization – We can compute and visualize p(attribute|feature vector), i.e., $p(x_n^d|\mathbf{z}_n, \mathbf{B}^d)$



Application to Clinical Trials

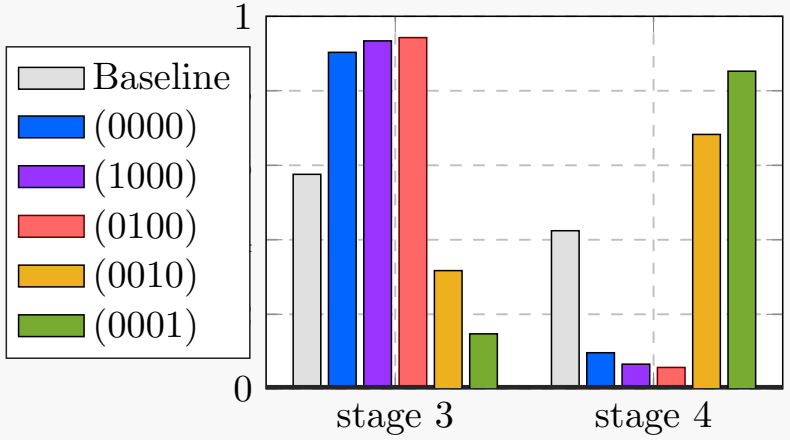
Goal: effects of a new drug for prostate cancer

Dataset contains 502 patients and 16 attributes, from which we make use of the following attributes:

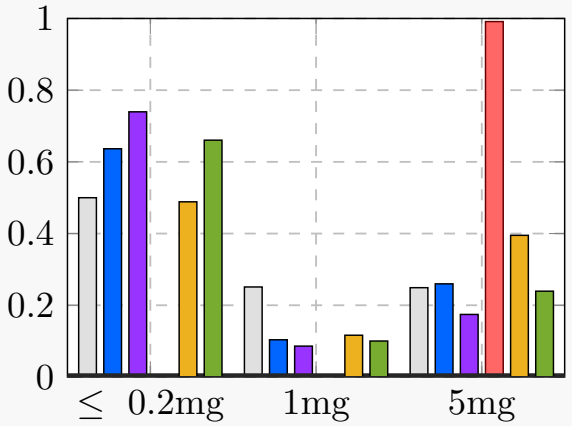
Attribute description	Type of variable
Stage of the cancer	Categorical with 2 categories
DES treatment level	Ordinal with 3 categories
Tumor size in cm^2	Count data
Serum Prostatic Acid Phosphatase (PAP)	Positive real-valued
Prognosis Status (outcome of the disease)	Categorical with 4 categories

Application to Clinical Trials

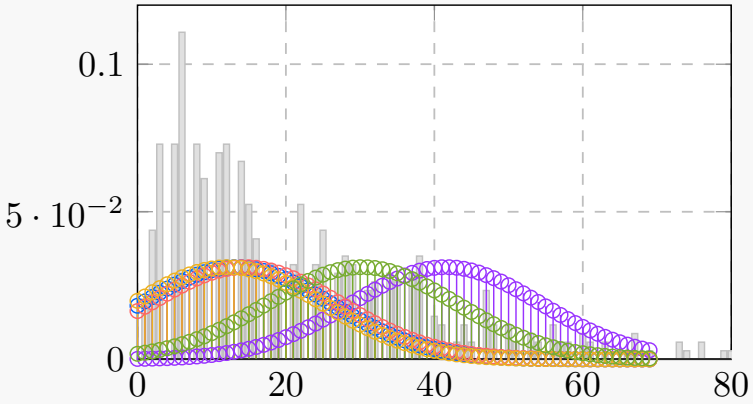
Feature patterns:



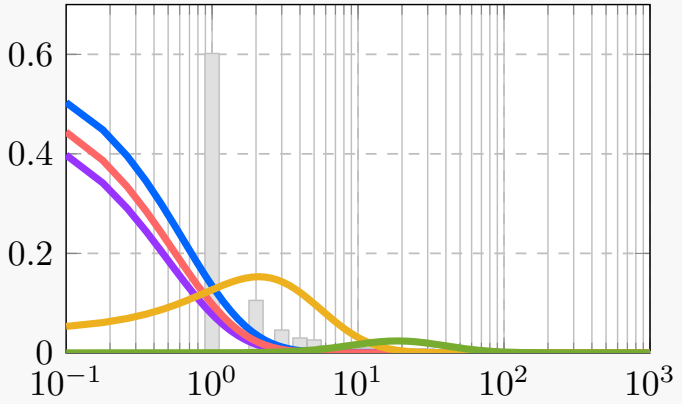
(a) Type of Cancer



(b) Drug Level



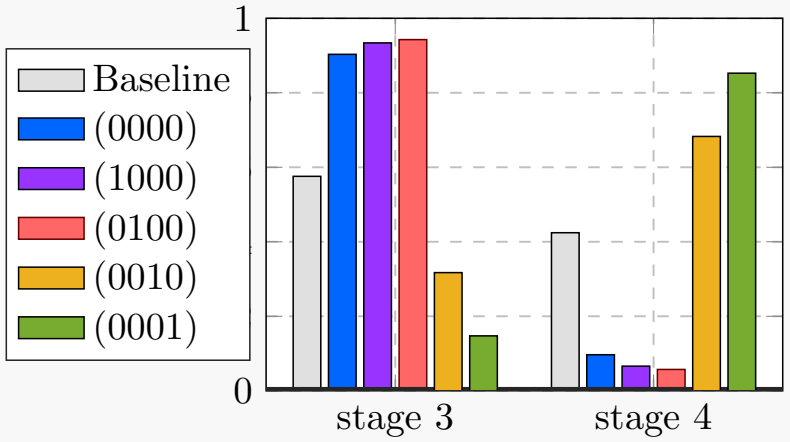
(c) Size of Primary Tumor (cm^2)



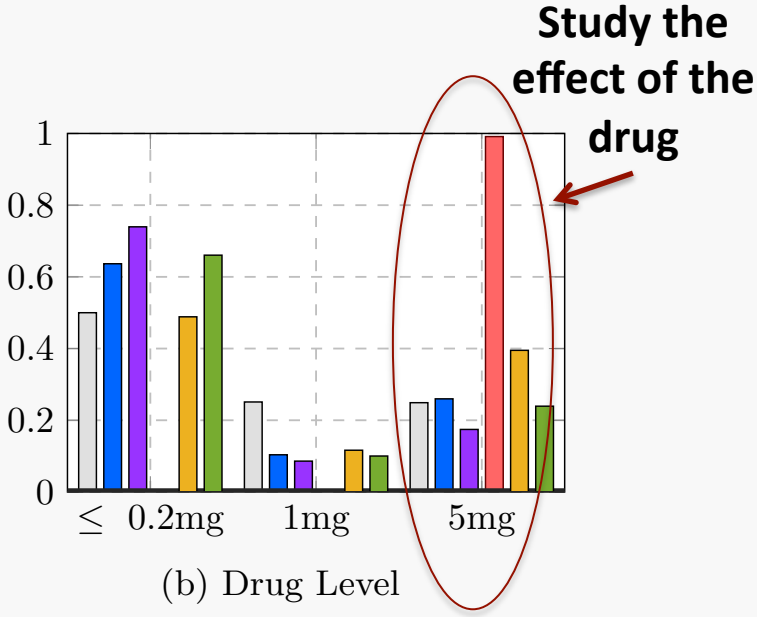
(d) Serum Prostatic Acid Phosphatase

Application to Clinical Trials

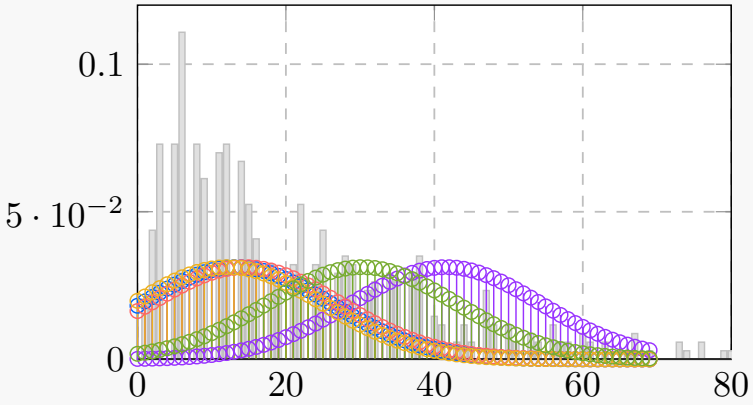
Feature patterns:



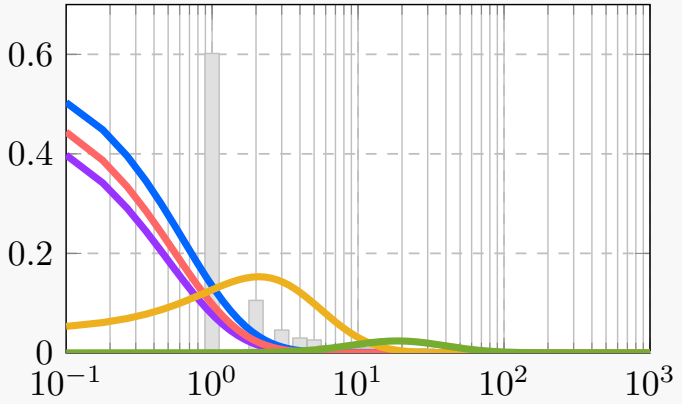
(a) Type of Cancer



(b) Drug Level



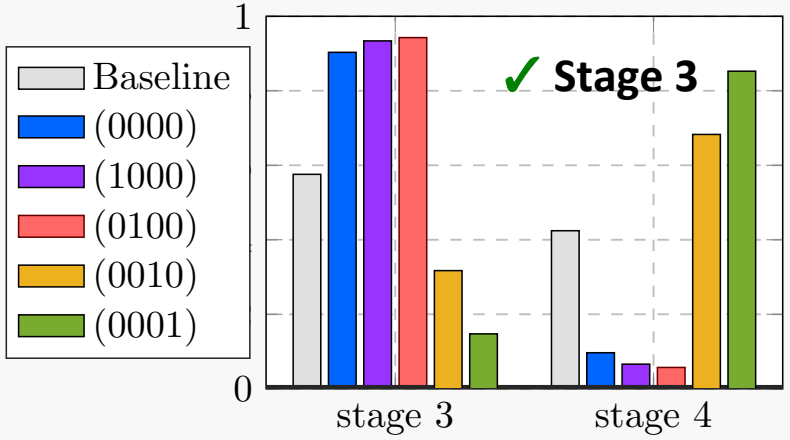
(c) Size of Primary Tumor (cm^2)



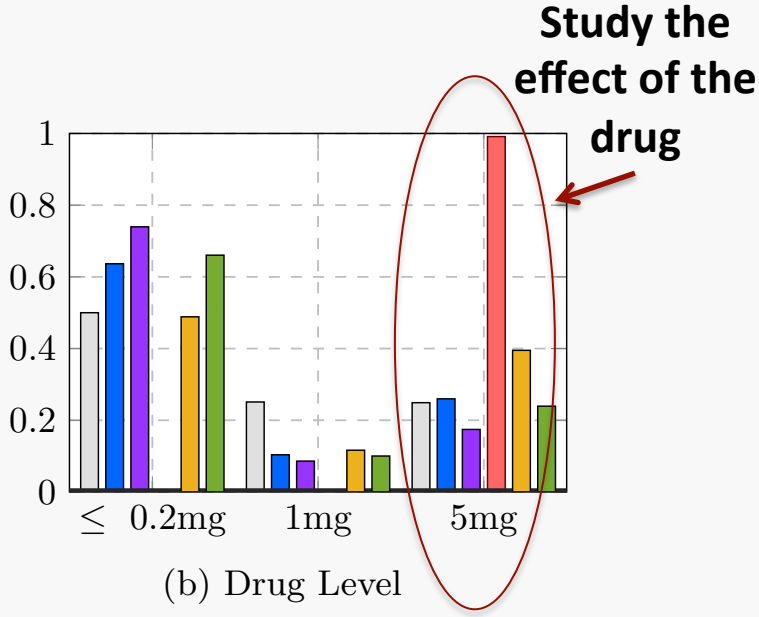
(d) Serum Prostatic Acid Phosphatase

Application to Clinical Trials

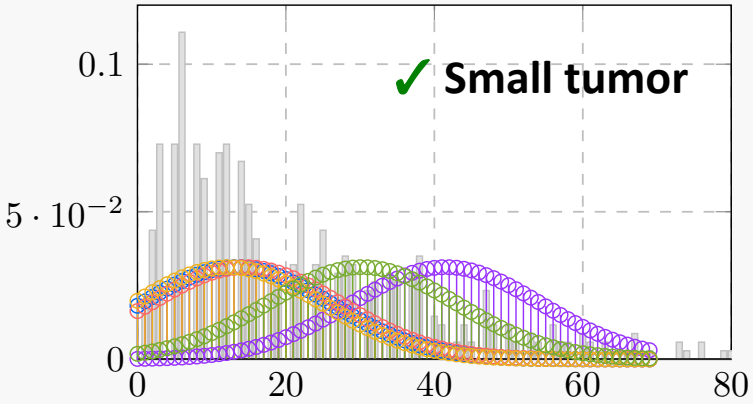
Feature patterns:



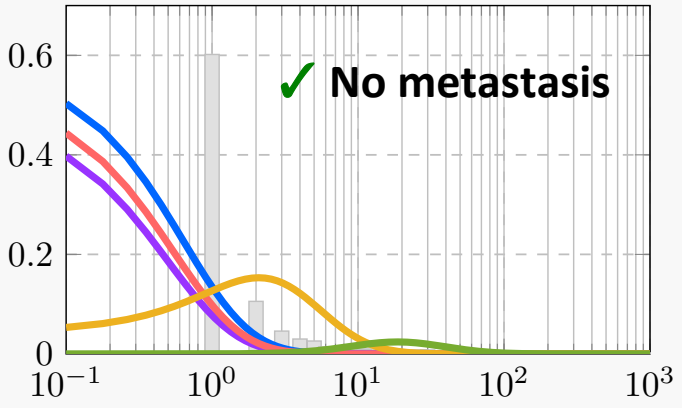
(a) Type of Cancer



(b) Drug Level



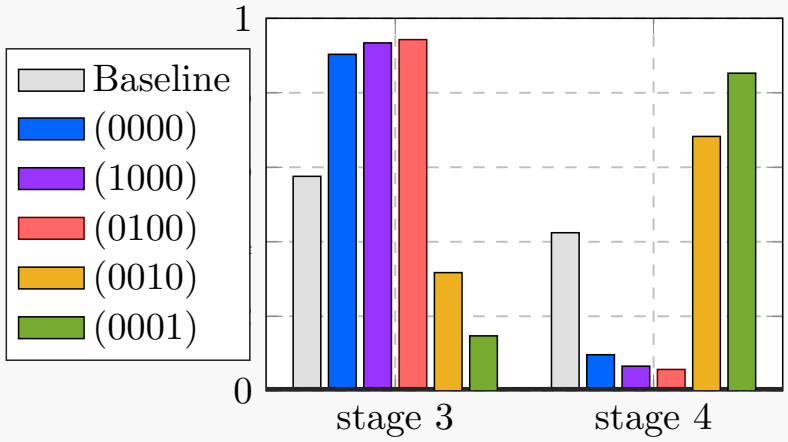
(c) Size of Primary Tumor (cm^2)



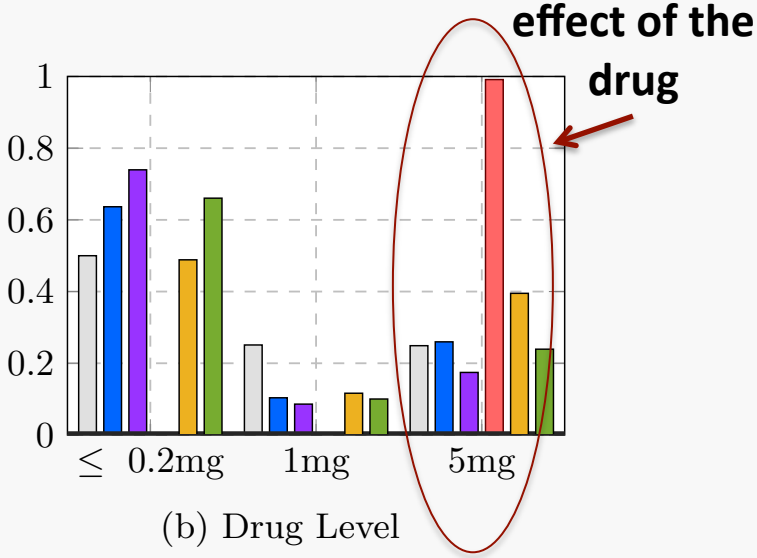
(d) Serum Prostatic Acid Phosphatase

Application to Clinical Trials

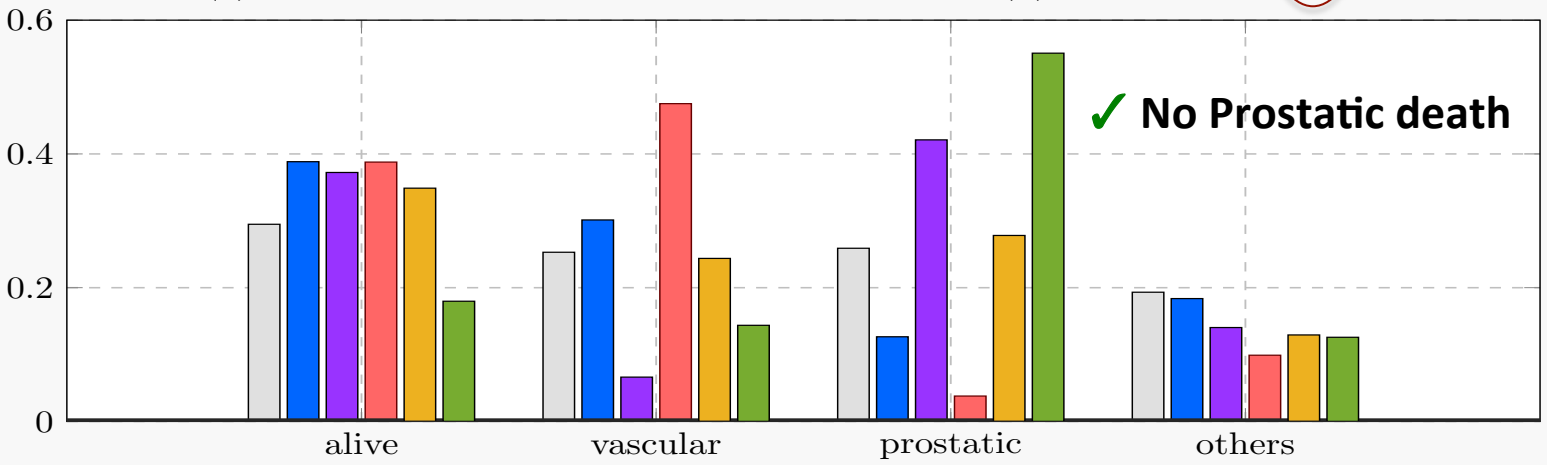
Feature patterns:



(a) Type of Cancer



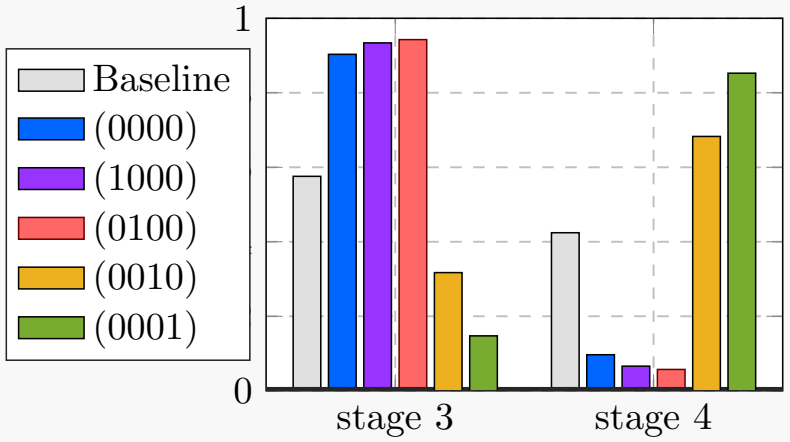
(b) Drug Level



(e) Prognosis Status

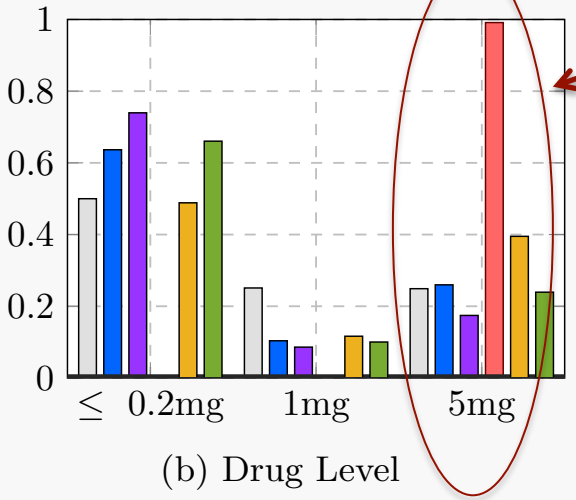
Application to Clinical Trials

Feature patterns:

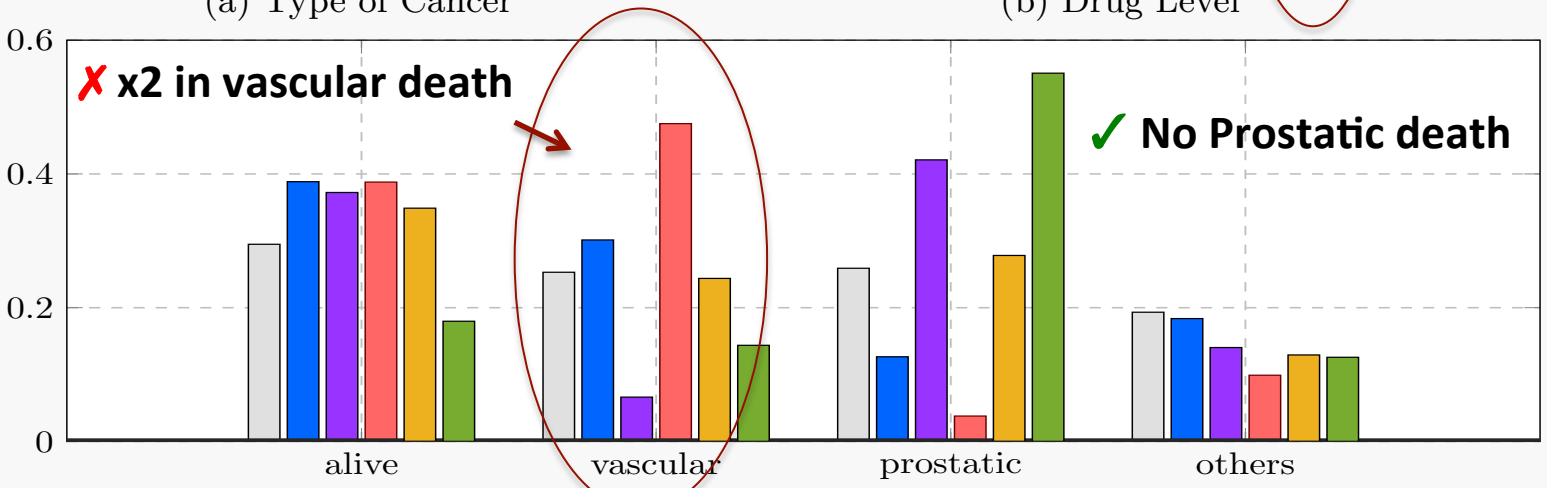


(a) Type of Cancer

Study the effect of the drug



(b) Drug Level



(e) Prognosis Status

Further results

1) Further details: https://arxiv.org/abs/1706.03779
   i. Other datasets: Psychiatry, electoral elections…
   ii. Other applications: Missing data estimation

2) Available Software: https://ivaleram.github.io/GLFM/
   i. User interfaces in Matlab, Python and R.
   ii. Available functions for:
      - Inference
      - Predictions
      - Visualization

Further results

1) Further details: https://arxiv.org/abs/1706.03779
   i. Other datasets: Psychiatry, electoral elections…
   ii. Other applications: Missing data estimation

2) Available Software: https://ivaleram.github.io/GLFM/
   i. User interfaces in Matlab, Python and R.
   ii. Available functions for:
      - Inference
      - Predictions
      - Visualization

Thanks!