CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

Learning Exceptional Subgroups by End-to-End Maximizing KL-divergence

Sacha Xu*, **Nils Philipp Walter***, Janis Kalofolias, Jilles Vreeken

*Equal contribution

Exploratory vs. Predictive ML

Exploratory ML

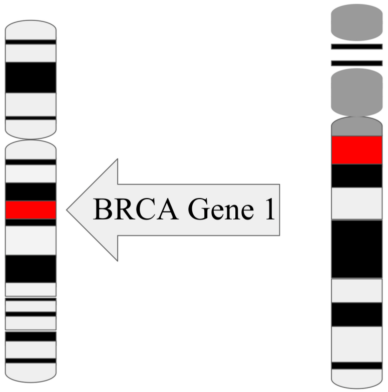
Exploratory + Predictive ML

Predictive ML

Best of both worlds

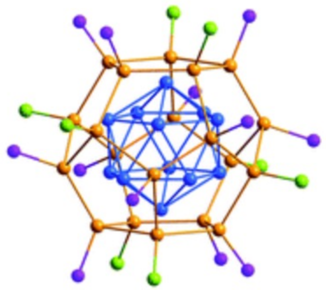
Examples

Breast Cancer

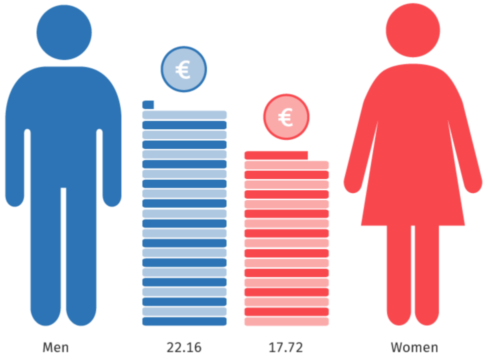


Malware Analysis

Materials Science



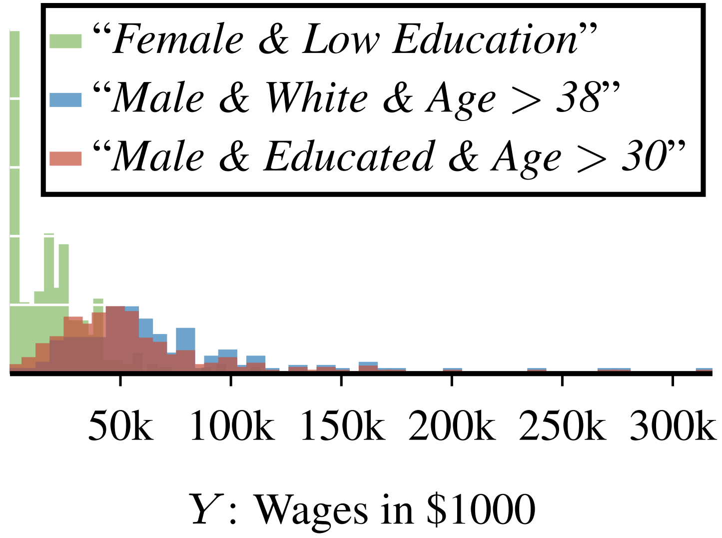
Census Data



Motivation – SyFlow

Census Data

Sex	Height	Race	Education	Age	Income
♀	168	White	12	72	17k
♂	163	White	11	55	23k
♀	160	White	**5**	62	**1k**
♂	188	White	16	38	63k
♀	165	White	**9**	45	**4k**
♂	172	White	12	78	71k
♀	180	White	**8**	74	**1k**



Task – Subgroup Discovery:

1. Find **exceptional** subgroups
2. With an **interpretable** description

Subgroup Discovery till now

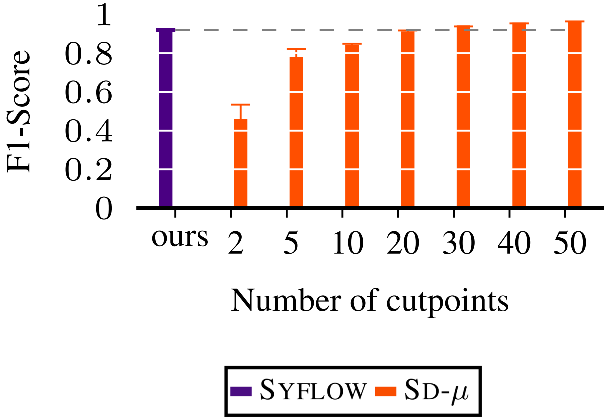
Prototypical Subgroup Discovery

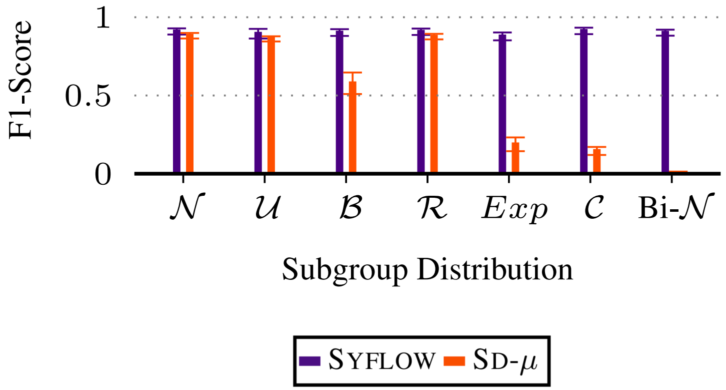
1. Generate boolean predicates
   i. Categorical: Sex=♀
   ii. Continuous: 170 < height < 180 ..
2. Use a (parametric) exceptionality measure
3. Combinatorially search the best subgroup

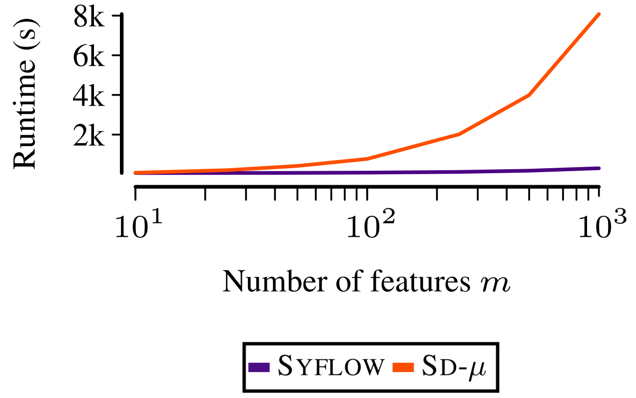
Sex	Height	Race	Education	Age	Income
♀	168	White	12	72	17k
♂	163	White	11	55	23k
♀	160	White	5	62	1k
♂	188	White	16	38	63k
♀	165	White	14	45	4k
♂	172	White	12	78	71k
♀	180	White	8	74	1k

Three major problems

1. Highly dependents on discretization
2. Only works for assumed distribution
3. Does not scale to large dimensions







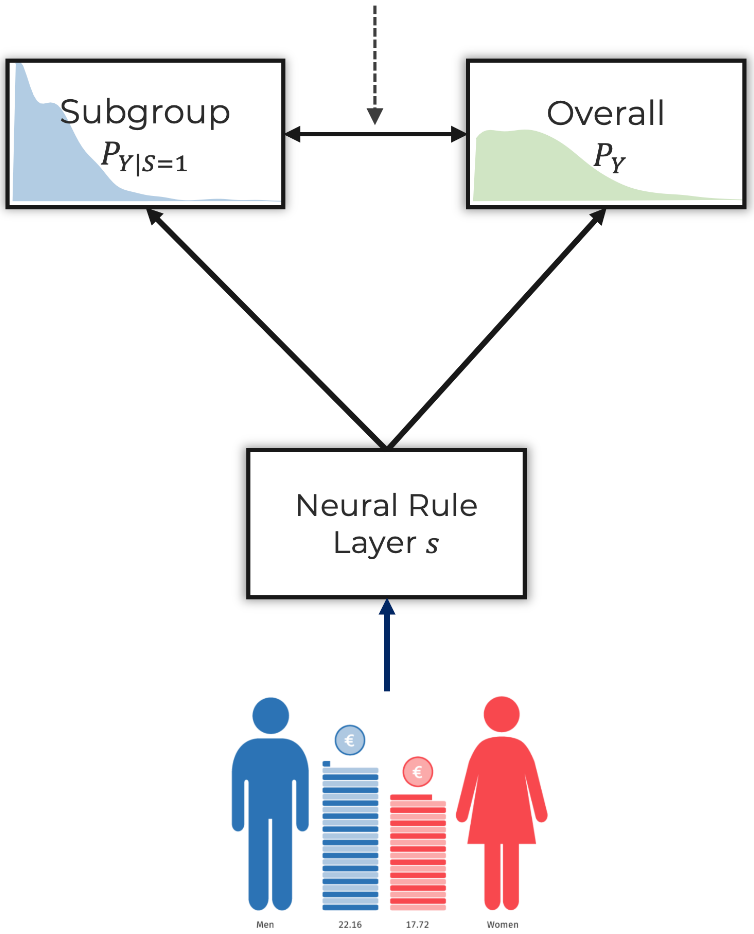
SyFlow – In a nutshell

Subgroup Discovery

1. Dependent on Pre-Discretization
2. Strong assumptions on the target distribution
3. Combinatorial optimization

1. Learn predicates end-to-end
   → **Accurate Discretization**
2. Use Normalizing Flows (NFs)
   → **No assumptions**
3. Continous optimization
   → **Highly scalable**

$$\max\ KL(P_{Y|S=1} \mid P_Y)$$



(Rezende & Mohamed. 2015)

SyFlow – Neural Rule Layer I

Goal: Find an crisp interpretable description

$$\sigma(x) = \neg Smoker \land 44 < Age < 64$$

Ingredients:

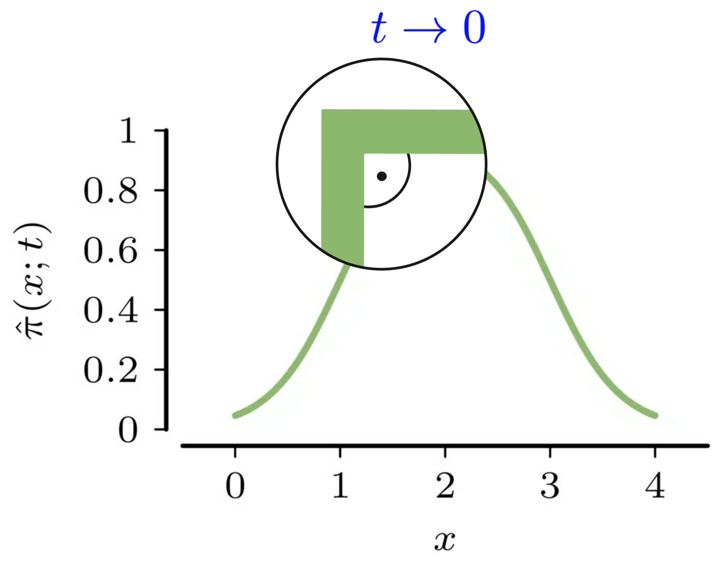
1. Differentiable binning predicate
2. [illegible]

$$\hat{\pi}(x_i; \alpha_i, \beta_i, t) = \frac{e^{\frac{1}{t}(2x_i - \alpha_i)}}{e^{\frac{1}{t}x_i} + e^{\frac{1}{t}(2x_i - \alpha_i)} + e^{\frac{1}{t}(3x_i - \alpha_i - \beta_i)}}$$

- Differentiable analog of:

$$\pi(x_i; \alpha_i, \beta_i) = \begin{cases} 1 & \text{if } \alpha_i < x_i < \beta_i \\ 0 & \text{otherwise} \end{cases}$$

- Temperature t controls crispness



Theorem 1 *Given its lower and upper bounds $\alpha_i, \beta_i \in \mathbb{R}$, the soft predicate of Eq. (1) applied on $x \in R$ converges to the crisp predicate that decides whether $x \in (\alpha, \beta)$,*

$$\lim_{t \to 0} \hat{\pi}(x_i; \alpha_i, \beta_i, t) = \begin{cases} 1 & \text{if } \alpha_i < x_i < \beta_i \\ 0.5 & \text{if } x_i = \alpha_i \lor x_i = \beta_i \\ 0 & \text{otherwise} \end{cases} .$$

(Yang et. al. 2018)

SyFlow – Neural Rule Layer II

Ingredients:

1. Differentiable binning predicate

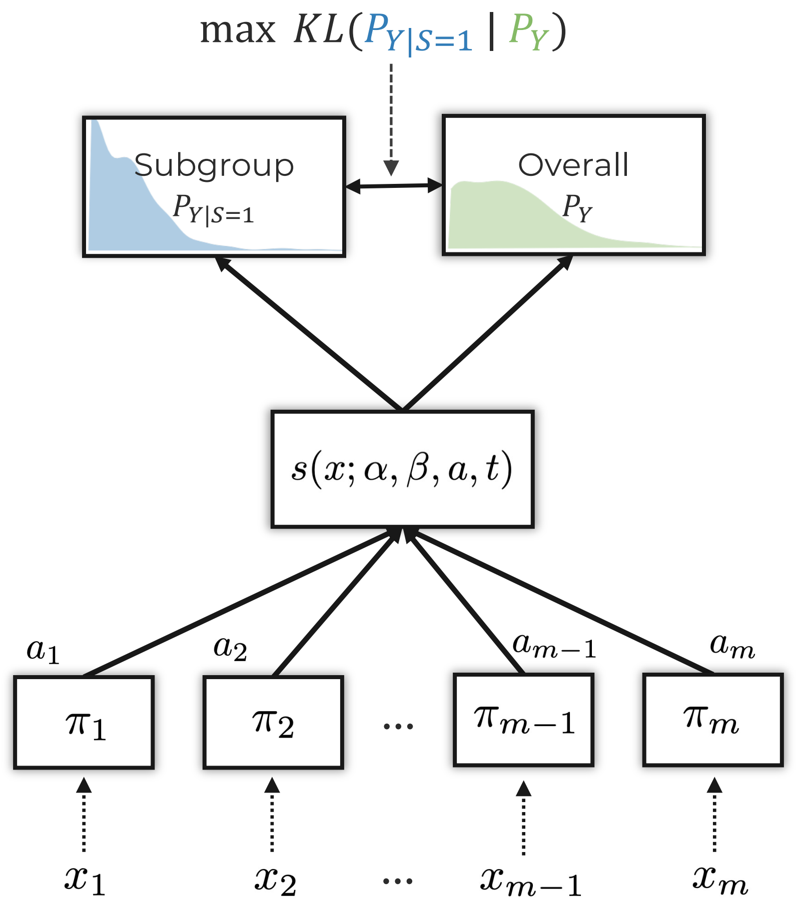
$$\hat{\pi}(x_i; \alpha_i, \beta_i, t) = \frac{e^{\frac{1}{t}(2x_i - \alpha_i)}}{e^{\frac{1}{t}x_i} + e^{\frac{1}{t}(2x_i - \alpha_i)} + e^{\frac{1}{t}(3x_i - \alpha_i - \beta_i)}}$$

2. Differentiable logical *AND*

$$\mathcal{M}(x) = \frac{m}{\sum_{i=1}^{m} \hat{\pi}(x_i; \alpha_i, \beta_i, t)^{-1}}$$

- Harmonic means behaves like an *AND*
  1. If one $\hat{\pi}(x_i; \alpha_i, \beta_i, t) = 0 \Rightarrow \mathcal{M}(x) = 0$
  2. If all $\hat{\pi}(x_i; \alpha_i, \beta_i, t) = 1 \Rightarrow \mathcal{M}(x) = 1$
- How to turn off useless predicates?

$$s(x; \alpha, \beta, a, t) = \frac{\sum_{i=1}^{m} a_i}{\sum_{i=1}^{m} a_i \hat{\pi}(x_i; \alpha_i, \beta_i, t)^{-1}}$$

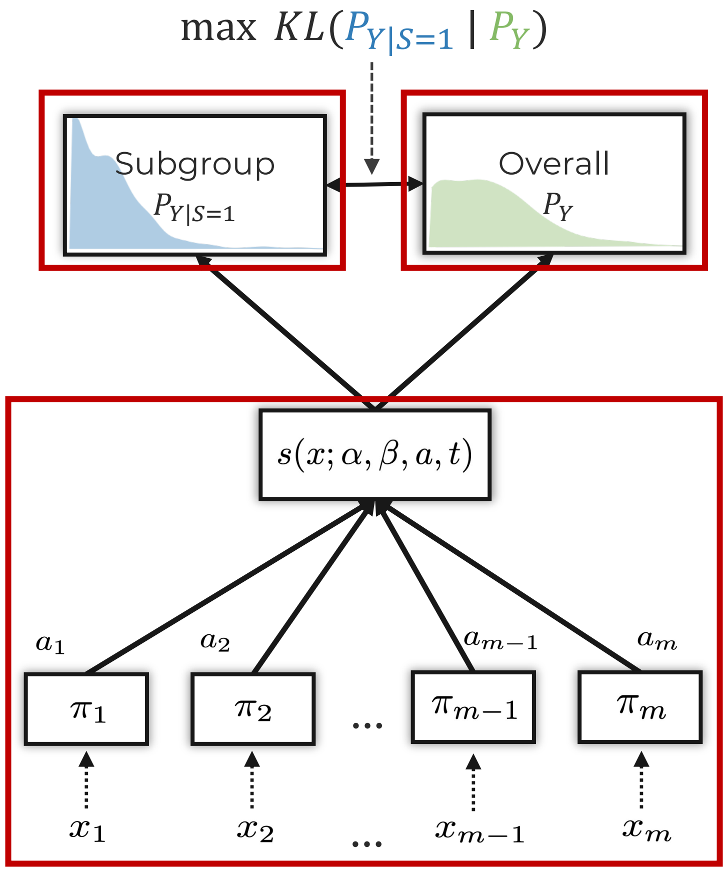


Fully differentiable!

SyFlow – Finding general & diverse subgroups

Our objective

$$D_{\mathrm{WKL}}\left(P_{Y|S=1} \| P_Y\right) = \left(\frac{n_s}{n}\right)^{\gamma} \hat{D}_{\mathrm{KL}}\left(P_{Y|S=1} \| P_Y\right)$$



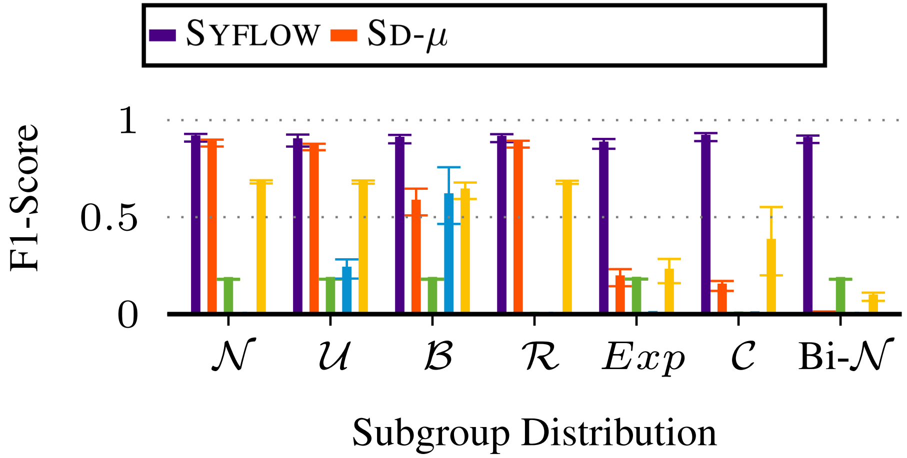
Optimization

1. Learn the overall distribution P_Y
2. Learn the subgroup distribution $P_{Y|S=1}$
3. Optimize classifier weights and bins
4. Output: Subgroup

(Steps 2–3: Repeat for N steps; steps 2–4: Repeat for k subgroups)

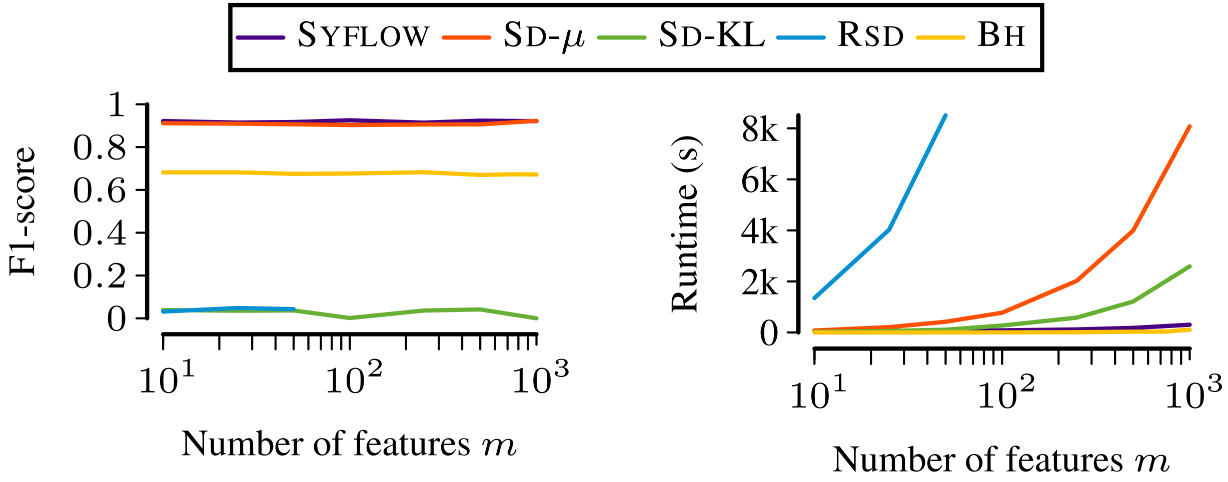
Experiments – Synthetic

Target distributions



SyFlow is robust to various target distributions.

Scalability in m



SyFlow finds a good balance between accuracy and runtime.

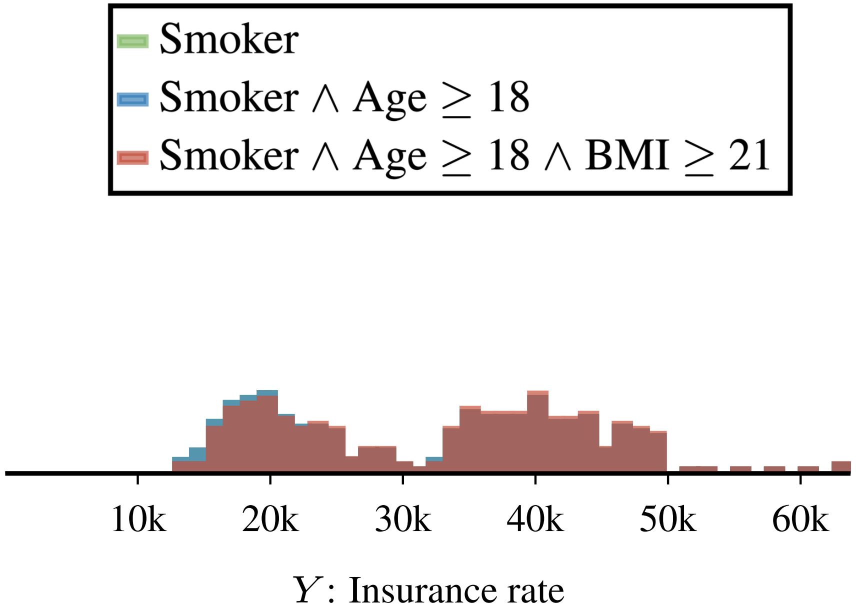
(Lemmerich & Becker, 2018; Lemmerich & Becker, 2018 + van Leeuwen & Knobbe, 2011; Proenca et al., 2022; Friedman & Fisher, 1999)

Experiments – Real World

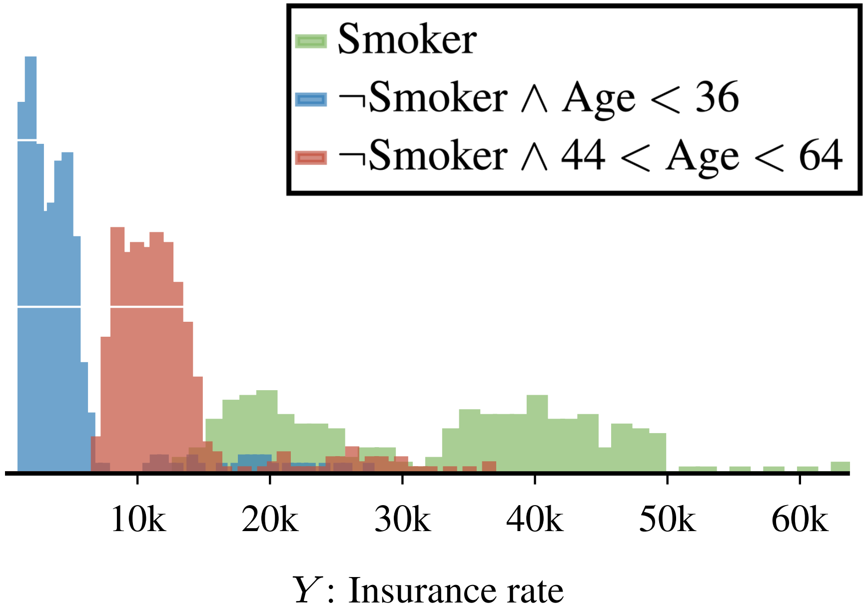
	D_{KL}					BC					AMD				
	ours	SD-KL	SD-μ	RSD	BH	*ours*	SD-KL	SD-μ	RSD	BH	*ours*	SD-KL	SD-μ	RSD	BH
Abalone	**0.14**	0.02	0.12	0	0.05	**0.66**	0.99	0.93	1	0.87	0.73	0.25	**0.84**	0	0.16
Airquality	0.22	0.22	**0.24**	0	0.0	**0.62**	0.86	0.79	1	1.0	0.37	**0.53**	0.49	0	0.0
Automobile	0.22	0.24	0.23	**0.26**	0.21	0.64	0.85	0.79	0.64	**0.6**	1838	**2807**	2683	2218	2475
Bike	**0.17**	0.1	0.15	0.17	0.13	**0.64**	0.95	0.9	0.67	0.73	584	570	**630**	431	622
California	**0.13**	0.06	0.11	0	0.0	**0.72**	0.97	0.93	1	1.0	0.25	0.3	**0.32**	0	0.0
Insurance	**0.27**	0.13	0.26	0	0.19	0.55	0.93	**0.52**	1	0.84	3845	**3973**	3845	0	1518
Mpg	**0.27**	0.26	0.24	0.21	0.24	0.57	0.76	0.8	**0.47**	0.61	**2.99**	2.85	2.96	1.66	2.79
Student	0.08	0.03	0.08	**0.09**	0.04	0.86	0.99	0.94	**0.71**	0.97	0.46	0.52	**0.69**	0.47	0.45
Wages	**0.1**	0.02	0.1	0	0.03	**0.81**	0.99	0.9	1	0.99	**6043**	2994	5916	0	5149
Wine	**0.08**	0.0	0.06	0	0.01	**0.89**	1.0	0.97	1	0.97	0.17	0.04	**0.19**	0	0.04
Avg. rank	**1.5**	3.5	2.1	3.5	3.6	**1.4**	4.0	2.8	3.3	2.9	2.6	2.4	**1.5**	4.5	3.6

Experiments – Insurance Dataset

SD-μ

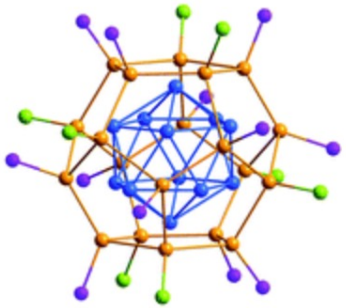


SyFlow



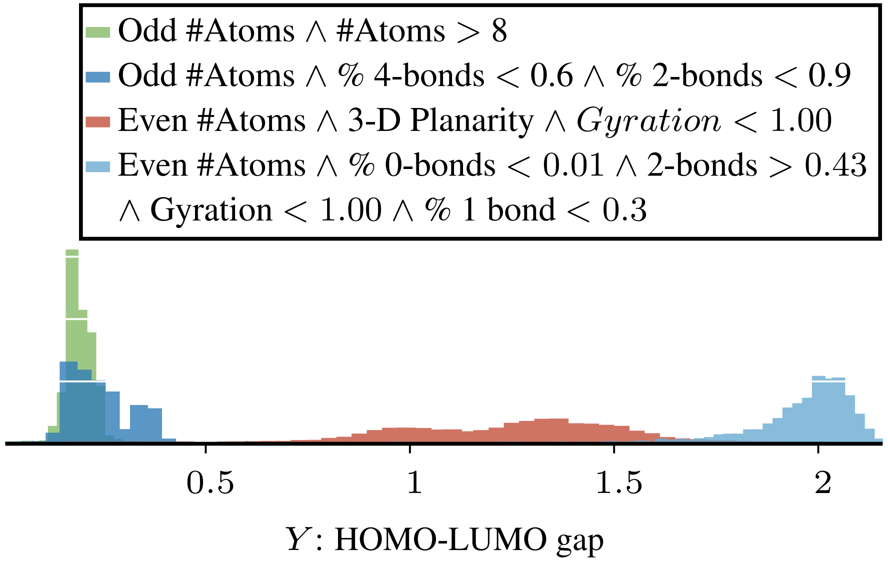
Experiments – Materials Sciences

Gold Nanoclusters



- Number of Atoms
- Even #Atoms
- 3-D Planarity

Target: HOMO-LUMO gap
~ stability and conductivity

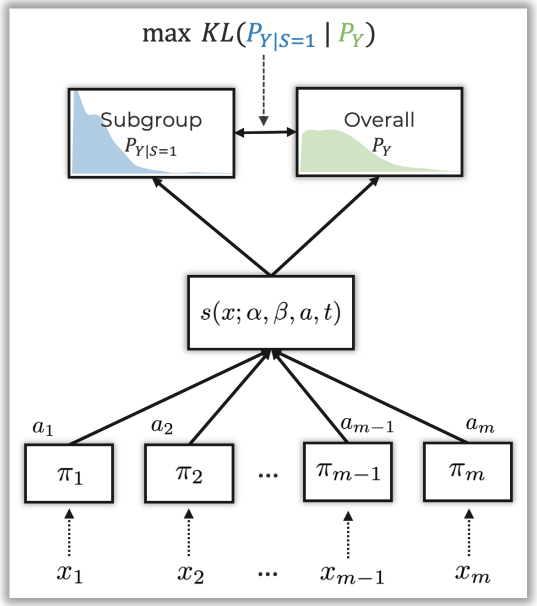


Conclusion

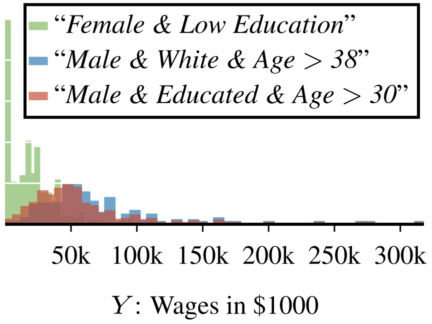
Census Data

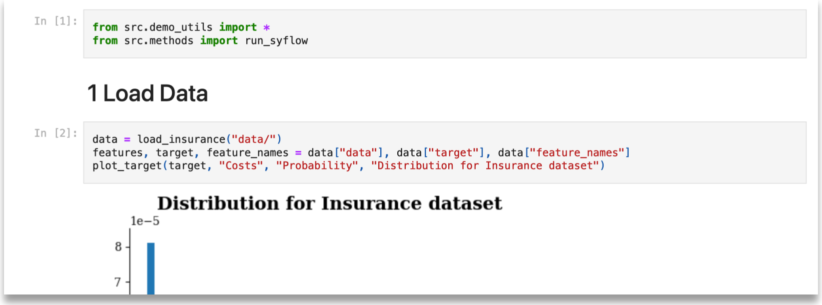
Sex	Height	Race	Edu.	Age	Income
♀	168	White	12	72	17k
♂	163	White	11	55	23k
♀	160	White	**5**	62	**1k**
♂	188	White	16	38	63k
♀	165	White	**9**	45	**4k**
♂	172	White	12	78	71k
♀	180	White	**8**	74	**1k**

SYFLOW



Discovered Subgroups





Paper

Code

References

[1] https://en.wikipedia.org/wiki/BRCA_mutation

[2] Walter, N. P., Fischer, J., & Vreeken, J. (2023). Finding Interpretable Class-Specific Patterns through Efficient Neural Search. *arXiv preprint arXiv:2312.04311*.

[3] Breiman, L. 1984. Classification and regression trees. Routledge

[4] Proenca, H. M.; and van Leeuwen, M. 2020. Interpretable multiclass classification by MDL-based rule lists. *Information Sciences.*

[5] Pellegrina, L.; Riondato, M.; and Vandin, F. 2019. SPuManTE: Significant pattern mining with unconditional testing. *In Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD*).

[6] Hedderich, M. A.; Fischer, J.; Klakow, D.; and Vreeken, J. 2022. Label-descriptive patterns and their application to characterizing classification errors. *In Proceedings of the International Conference on Machine Learning (ICML).*

[7] Wang, Z.; Zhang, W.; Liu, N.; and Wang, J. 2021. Scalable rule-based representation learning for interpretable classification. *In Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS).*

[8] Ulianova, S. 2017. Cardiovascular Disease dataset. *https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset.*

[9] Patil, P.; and Rathod, P. 2020. Disease Symptom Prediction. *https://www.kaggle.com/datasets/itachi9604/disease-symptom-description-dataset*

[10] The Cancer Genome Atlas (TCGA). *https://www.cancer.gov/tcga*.

[11] The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature.*

[12] Rezende, D., & Mohamed, S. 2015. Variational inference with normalizing flows. *In Proceedings of the International Conference on Machine Learning (ICML).*

References

Images on slide 10

1. https://en.wikipedia.org/wiki/BRCA_mutation
2. https://benchmarks.elsa-ai.eu/
3. Kenzler, S., & Schnepf, A. (2021). Metalloid gold clusters–past, current and future aspects. *Chemical Science*.
4. https://www.destatis.de/DE/Presse/Pressemitteilungen/2020/03/PD20_097_621.html

Images on slide 11

1. Böhle, M., Fritz, M., & Schiele, B. (2022). B-cos networks: Alignment is all we need for interpretability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
2. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., & Riedmiller, M.A. (2013). Playing Atari with Deep Reinforcement Learning. *ArXiv, abs/1312.5602*.

SyFlow – Objective

Approximating KL-Divergence

$$
\begin{aligned}
D_{\mathrm{KL}}\left(P_{Y|S=1} \| P_Y\right) &= \int_{y \in \mathcal{Y}} p_{Y|S=1}(y) \log\left(\frac{p_{Y|S=1}(y)}{p_Y(y)}\right) dy \\
&\approx \int_{y \in \mathcal{Y}} \int_{\mathbf{x} \in \mathbb{R}^m_{\in}} p_{Y,\mathbf{x}}(y, \mathbf{x}) \frac{p_{S=1|\mathbf{x}}(\mathbf{x})}{\mathbb{P}(S=1)} dx \log\left(\frac{p_{Y|S=1}(y)}{p_Y(y)}\right) dy \\
&\approx \frac{1}{n_s} \sum_{k=1}^{n} s(\mathbf{x}^{(k)}) \log\left(\frac{p_{Y|S=1}(y^{(k)})}{p_Y(y^{(k)})}\right)
\end{aligned}
$$

Objective for general & diverse subgroups

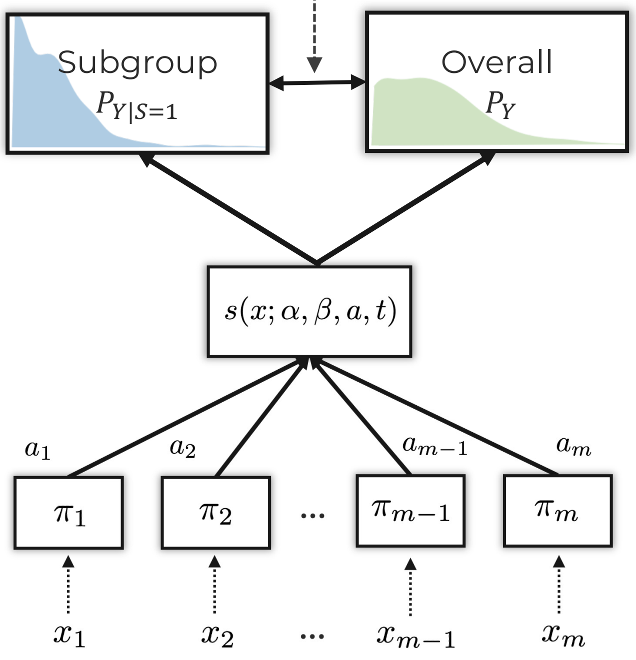
$$
\begin{aligned}
D_{\mathrm{WKL}}\left(P_{Y|S=1} \| P_Y\right) = &\left(\frac{n_s}{n}\right)^{\gamma} \hat{D}_{\mathrm{KL}}\left(P_{Y|S=1} \| P_Y\right) \\
&+ \lambda \sum_{j=1}^{j} \hat{D}_{\mathrm{KL}}\left(P_{Y|S=1} \| P_{Y|S_j=1}\right)
\end{aligned}
$$

→ Trade-off size and exceptionality

→ Diverse subgroups

SYFLOW

$$\max\ KL(P_{Y|S=1} \mid P_Y)$$

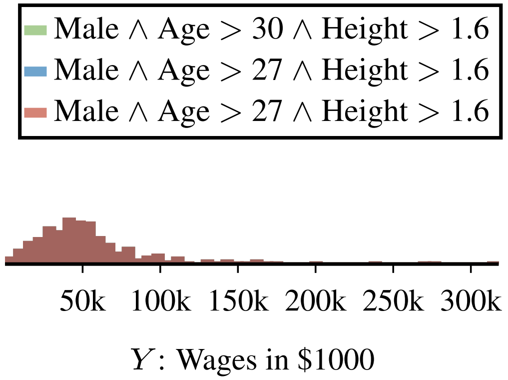


Fully differentiable!

Key contributions

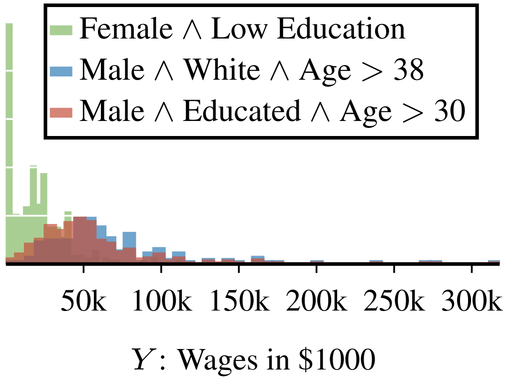
1. Continuous optimization to maximize KL-divergence
2. Normalizing Flows to accurately learn target distributions
3. Neuro-symbolic rule layer to learn interpretable subgroup descriptions

Traditional subgroup discovery



- Highly redundant
- Depends on pre-discretization
- Slow for large #features

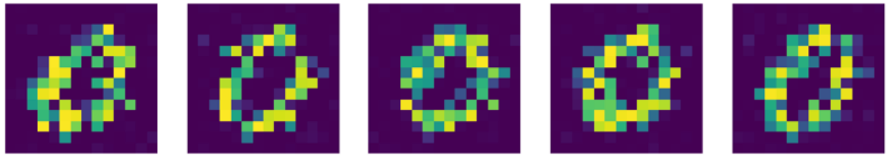
SYFLOW



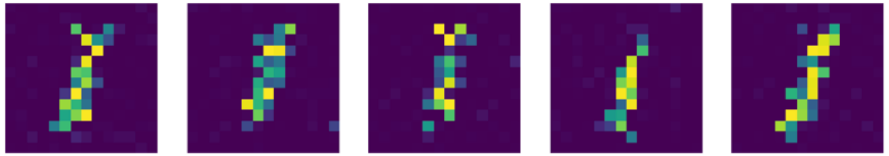
- Diverse set of subgroups
- Learns best discretization
- Highly scalable

On images

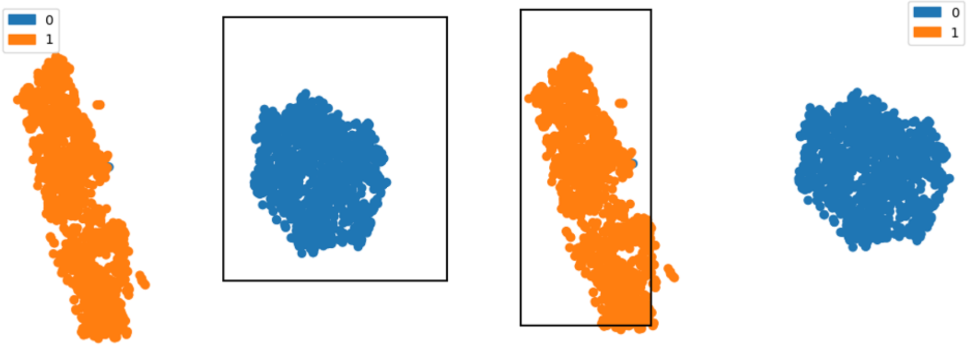
Distribution 0: $P_{Y|S_0=1}$



Distribution 1: $P_{Y|S_1=1}$



Rule 0: $s_0(X)$ Rule 1: $s_1(X)$



SyFlow – Table with bold numbers

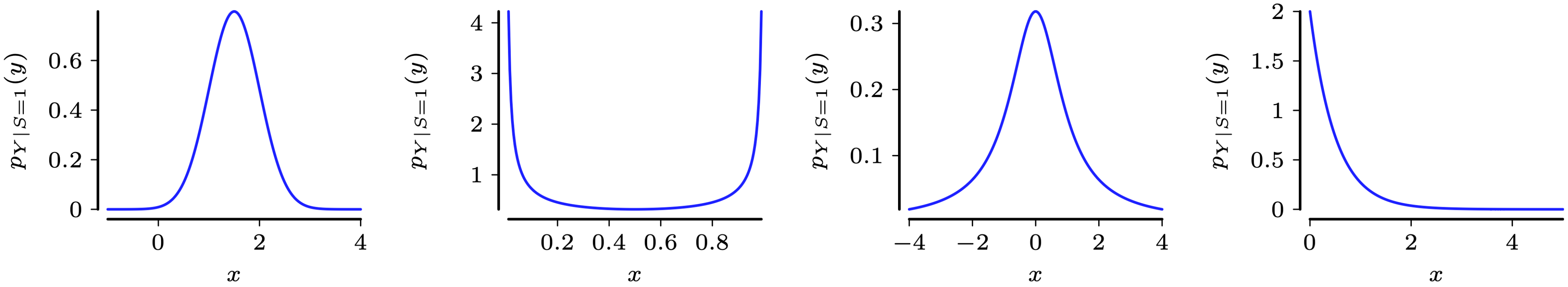
$$D_{KL}(P_{Y|S=1}, P_Y) = \sum_{y \in \mathcal{Y}} p_{Y|S=1}(y) \log(\frac{p_{Y|S=1}(y)}{p_Y(y)})$$

$$BC(P_{Y|S=1}, P_Y) = \sum_{y \in \mathcal{Y}} \sqrt{p_{Y|S=1}(y) p_Y(y)}$$

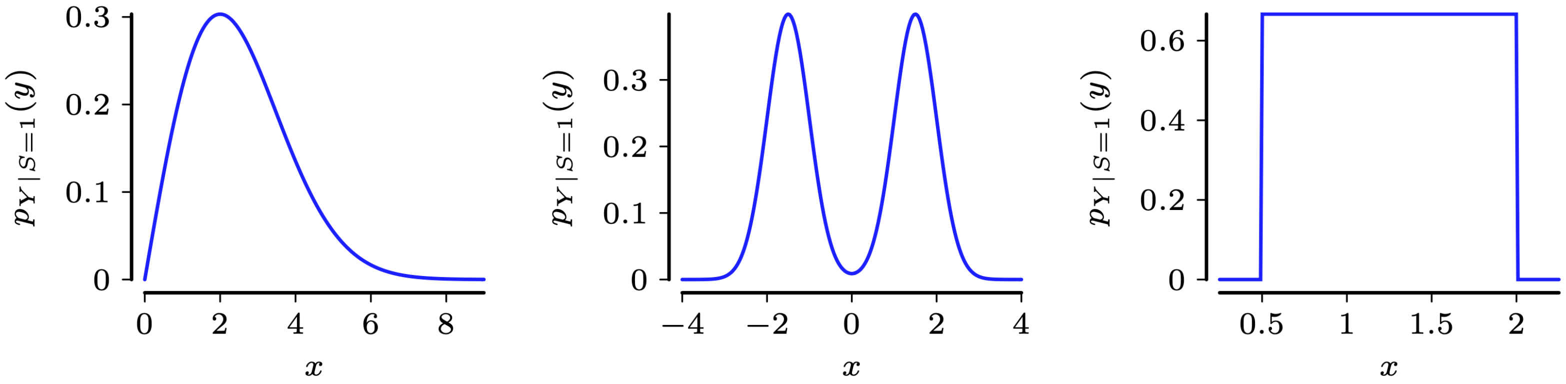
$$AMD(\mathcal{Y}_s, \mathcal{Y}) = \left| \left(\frac{1}{|\mathcal{Y}_s|} \sum_{y \in \mathcal{Y}_s} y \right) - \left(\frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} y \right) \right|$$

	D_{KL}					BC					AMD				
	ours	SD-KL	SD-μ	RSD	BH	*ours*	SD-KL	SD-μ	RSD	BH	*ours*	SD-KL	SD-μ	RSD	BH
Abalone	**0.14**	0.02	0.12	0	0.05	**0.66**	0.99	0.93	1	0.87	0.73	0.25	**0.84**	0	0.16
Airquality	0.22	0.22	**0.24**	0	0.0	**0.62**	0.86	0.79	1	1.0	0.37	**0.53**	0.49	0	0.0
Automobile	0.22	0.24	0.23	**0.26**	0.21	0.64	0.85	0.79	0.64	**0.6**	1838	**2807**	2683	2218	2475
Bike	**0.17**	0.1	0.15	0.17	0.13	**0.64**	0.95	0.9	0.67	0.73	584	570	**630**	431	622
California	**0.13**	0.06	0.11	0	0.0	**0.72**	0.97	0.93	1	1.0	0.25	0.3	**0.32**	0	0.0
Insurance	**0.27**	0.13	0.26	0	0.19	0.55	0.93	**0.52**	1	0.84	3845	**3973**	3845	0	1518
Mpg	**0.27**	0.26	0.24	0.21	0.24	0.57	0.76	0.8	**0.47**	0.61	**2.99**	2.85	2.96	1.66	2.79
Student	0.08	0.03	0.08	**0.09**	0.04	0.86	0.99	0.94	**0.71**	0.97	0.46	0.52	**0.69**	0.47	0.45
Wages	**0.1**	0.02	0.1	0	0.03	**0.81**	0.99	0.9	1	0.99	**6043**	2994	5916	0	5149
Wine	**0.08**	0.0	0.06	0	0.01	**0.89**	1.0	0.97	1	0.97	0.17	0.04	**0.19**	0	0.04
Avg. rank	**1.5**	3.5	2.1	3.5	3.6	**1.4**	4.0	2.8	3.3	2.9	2.6	2.4	**1.5**	4.5	3.6

SyFlow – Different Target distributions



(a) $\mathcal{N}(1.5, 0.5)$ (b) $\mathcal{B}(0.2, 0.2)$ (c) $\mathcal{C}(0, 1)$ (d) $Exp(0.5)$



(e) $\mathcal{R}(2)$ (f) $\mathcal{N}(-1.5, 0.5) + \mathcal{N}(1.5, 0.5)$ (g) $\mathcal{U}(0.5, 1.5)$