Automatic Bayesian Density Analysis

Antonio Vergari
Max-Plank-Institute IS

Alejandro Molina
TU Darmstadt

Robert Peharz
University of Cambridge

Zoubin Ghahramani
University of Cambridge
Uber AI Labs

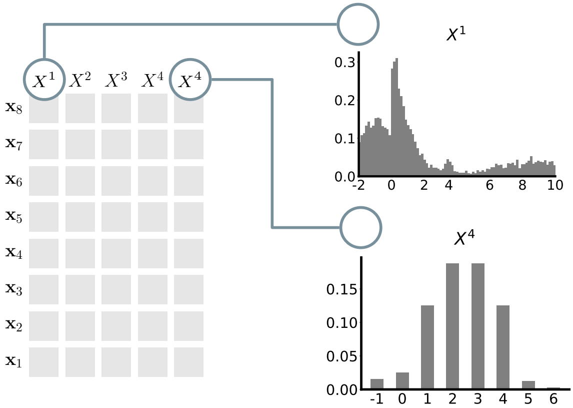
Kristian Kersting
TU Darmstadt

Isabel Valera
Max-Plank-Institute IS

31st January 2019 - ***AAAI19*** **Honolulu**

	X^1	X^2	X^3	X^4	X^5
$\mathbf{x}_8$					
$\mathbf{x}_7$					
$\mathbf{x}_6$					
$\mathbf{x}_5$					
$\mathbf{x}_4$					
$\mathbf{x}_3$					
$\mathbf{x}_2$					
$\mathbf{x}_1$					

Exploratory data analysis



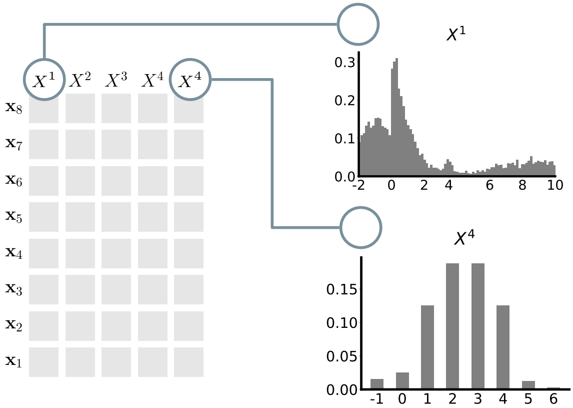
The “***data understanding***” pipeline

1a. ***infer statistical types***

$\Longrightarrow$ *is it a Gaussian?*

$\Longrightarrow$ *is it an ordinal RV?*

Exploratory data analysis



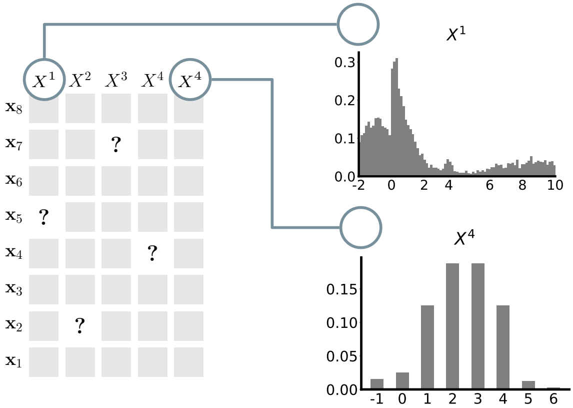
The "***data understanding***" pipeline

1a. ***infer statistical types***

1b. ***infer dependencies***

$\Rightarrow$ *does X^4 depend on X^1?*

Exploratory data analysis



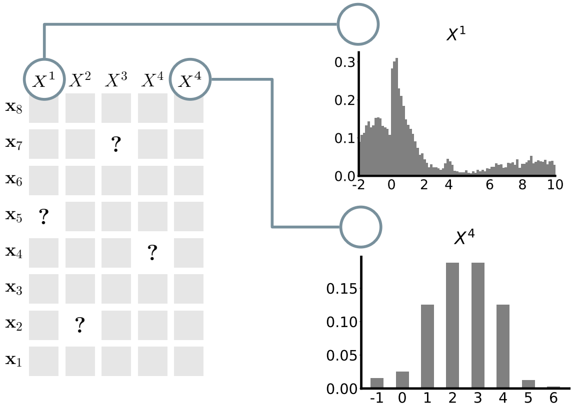
The "***data understanding***" pipeline

1a. ***infer statistical types***

1b. ***infer dependencies***

2. ***missing values estimation***

Exploratory data analysis



The "***data understanding***" pipeline

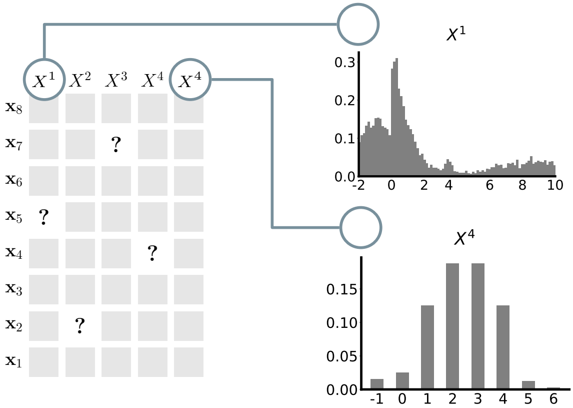
1a. ***infer statistical types***

1b. ***infer dependencies***

2. ***missing values estimation***

3. ***anomaly detection***

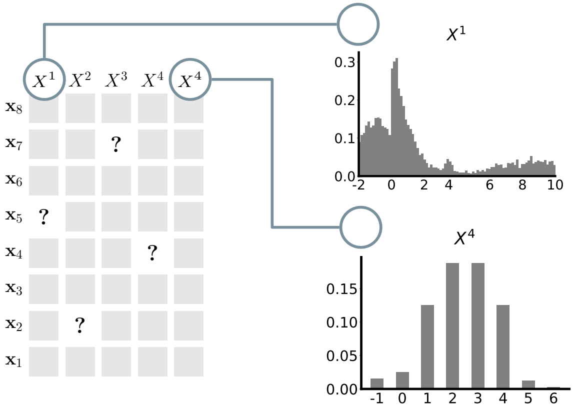
Exploratory data analysis



The “***data understanding***” pipeline

- **1a.** ***infer statistical types***
- **1b.** ***infer dependencies***
- **2.** ***missing values estimation***
- **3.** ***anomaly detection***
- **4.** ***pattern mining***

Exploratory data analysis



The "***data understanding***" pipeline

1a. ***infer statistical types***

1b. ***infer dependencies***

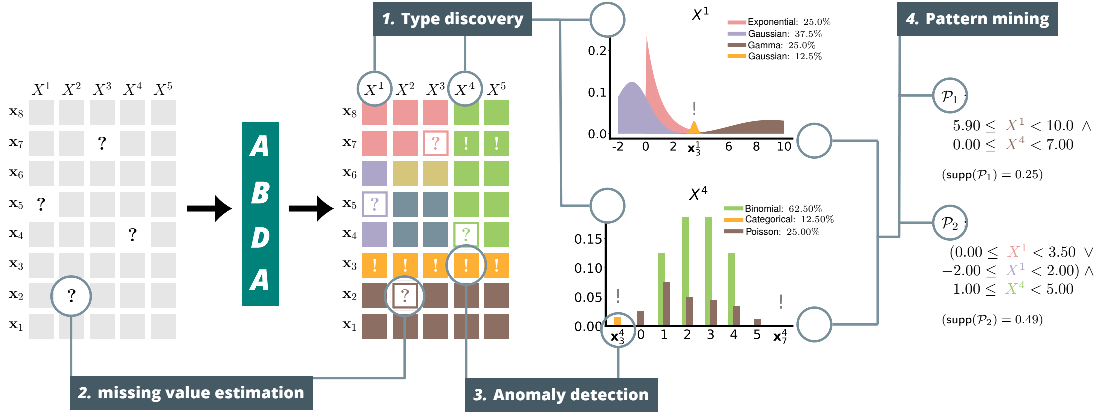
2. ***missing values estimation***

3. ***anomaly detection***

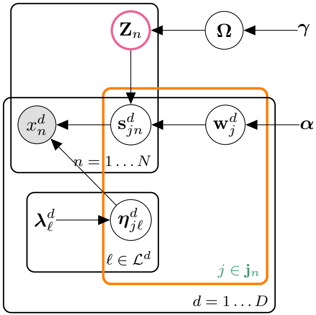
4. ***pattern mining***

...what if there is ***no statistician around*** ?

Automatic exploratory data analysis for non-statisticians



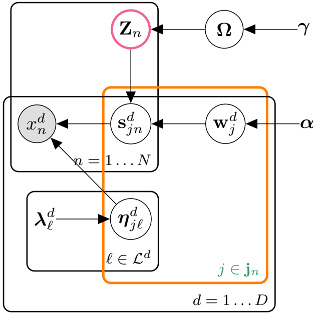
ABDA: representation

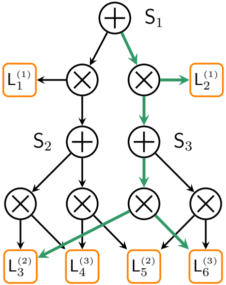


global level $\Rightarrow$ *dependencies among features*

local level $\Rightarrow$ *heterogeneous data types*

ABDA: representation

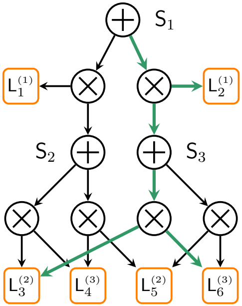




global level comprises a ***hierarchy of latent variables***

$\Rightarrow$ ***Sum-Product Network*** (SPN)

SPNs



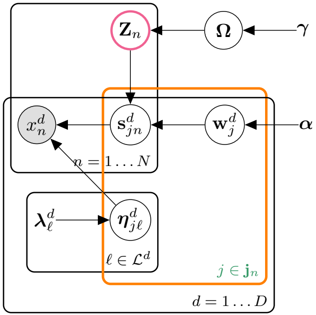
An SPN compiles a joint distribution into:

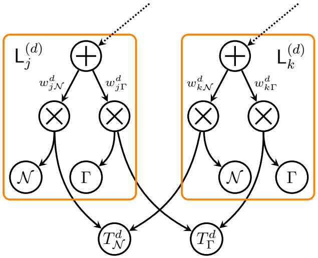
sum nodes $\Longrightarrow$ *mixtures*
product nodes $\Longrightarrow$ *independent groups of RVs*
leaf nodes $\Longrightarrow$ *small, tractable models*

SPNs allow to tractably

marginalize any set of RVs
perform **MAP inference** (approximate)

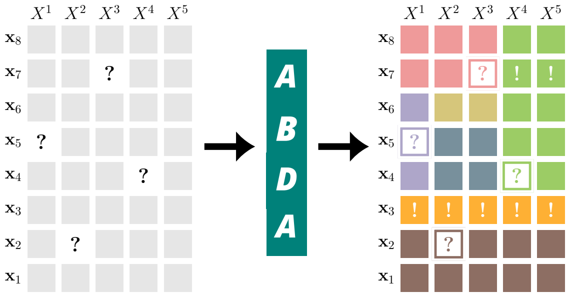
ABDA: representation





The ***local level*** comprises mixtures over ***collections of likelihood models***

ABDA: learning and inference



Learning the hierarchy of $\mathbf{Z}_n$

$\Rightarrow$ **data-agnostic structure learning**

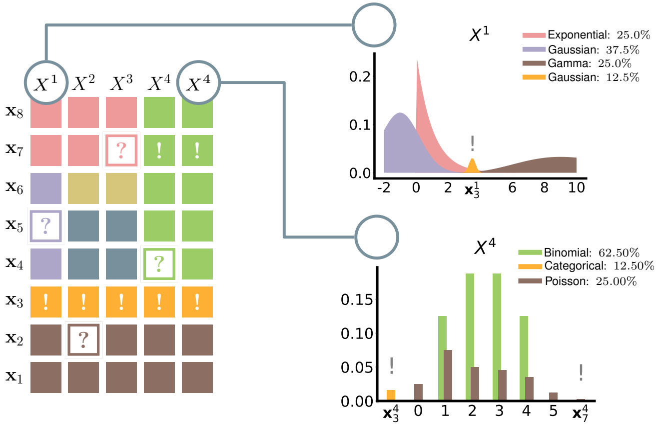
LVs partitioning the data

$\Rightarrow$ ***hierarchical co-clustering!***

Bayesian posterior inference

$\Rightarrow$ ***efficient Gibbs sampling***

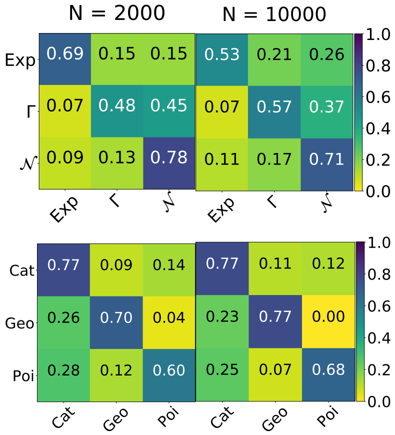
Type inference



Best likelihood models explaining the data

Mixture proportions represent ***uncertainty over likelihood models***

Type inference



Extensive ***synthetic experiments***

$\Longrightarrow$ *ground truth available*

Recovering uncertainty over likelihoods

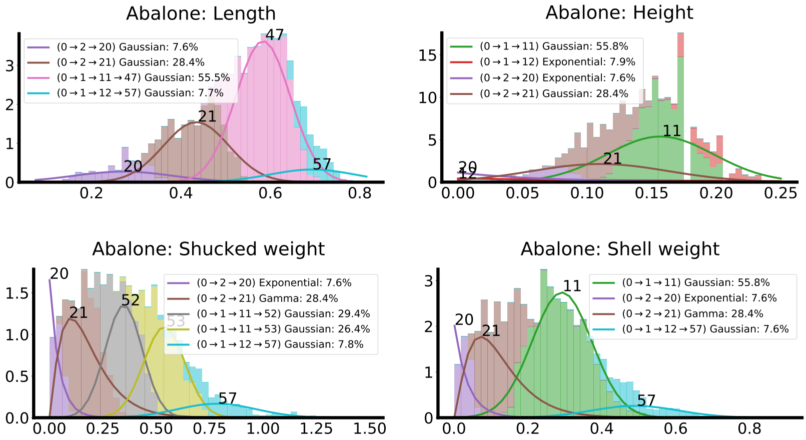
$\Longrightarrow$ *high cosine similarity over weight vectors*

Predicting most likely "**global types**"

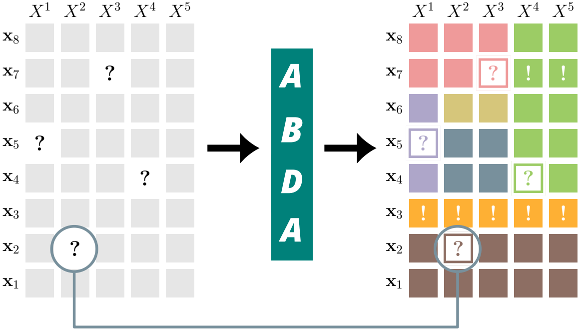
$\Longrightarrow$ *e.g., X^1 is Gaussian, X^4 is Poisson*

$\Longrightarrow$ *accurate up to small finite sample sizes*

Type inference



Missing values estimation



Marginalizing over missing values

$\Rightarrow$ *easy via SPN inference!*

Imputation via MAP inference

$\Rightarrow$ ***linear time*** *approximation*

$$\tilde{\mathbf{x}}_n^m = \operatorname*{argmax}_{\mathbf{x}_n^m} \mathcal{S}(\mathbf{x}_n^m \mid \mathbf{x}_n^o)$$

Missing values estimation

	50% missing		
	ISLV	**ABDA**	**MSPN**
Abalone	-0.89±0.36	-0.05±0.02	**0.14**
Adult	-	**-0.69±0.01**	-5.83
Austral.	-9.37±0.69	**-1.63±0.04**	-3.76
Autism	-2.67±0.16	**-1.24±0.01**	-1.57
Breast	-4.29±0.17	**-2.85±0.01**	-3.06
Chess	-2.58±0.04	**-1.87±0.01**	-3.92
Crx	-11.96±1.01	**-1.20±0.04**	-3.51
Derma.	-3.57±0.32	**-0.99±0.01**	-1.01
Diabetes	-12.52±0.52	**-2.37±0.09**	-4.01
German	-4.06±0.28	**-1.55±0.01**	-1.60
Student	-3.80±0.29	**-1.57±0.01**	-1.58
Wine	-1.34±0.01	-0.92±0.01	**-0.41**

Real-world benchmarks from UCI

$\Longrightarrow$ *repeated trials, different missing percentages*

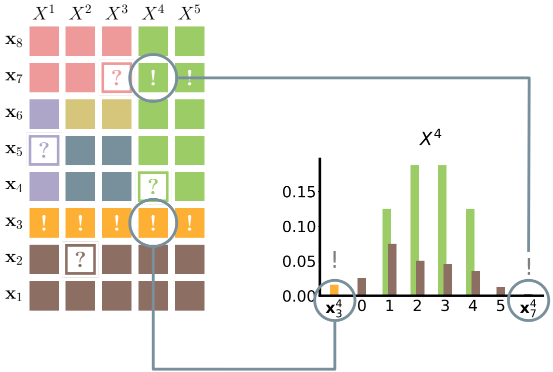
Faster and ***more accurate*** than ISLV

$\Longrightarrow$ *also modeling statistical type uncertainty*

and ***more robust*** than MSPNs

$\Longrightarrow$ *same data-agnostic structure learning*

Anomaly detection



Grouping anomalies together

$\Longrightarrow$ *micro-clusters*

relegating them into ***low-density regions***

$\Longrightarrow$ *e.g., tails of distributions*

log-likelihood as anomaly score

$\Longrightarrow$ *for **outliers** (transductive case)*

$\Longrightarrow$ *for **novelties** (inductive case)*

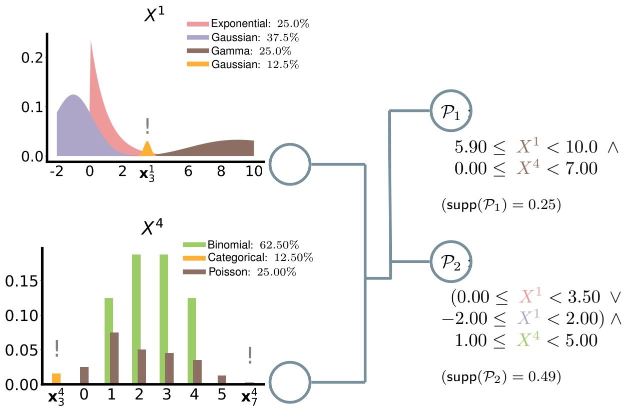
Anomaly detection

	1SVM	LOF	HBOS	ABDA
Aloi	51.71$\pm$0.02	**74.19**$\pm$0.70	52.86$\pm$0.53	47.20$\pm$0.02
Thyroid	46.18$\pm$0.39	62.38$\pm$1.04	62.77$\pm$3.69	**84.88**$\pm$0.96
Breast	45.77$\pm$11.1	98.06$\pm$0.70	94.47$\pm$0.79	**98.36**$\pm$0.07
Kdd99	53.40$\pm$3.63	46.39$\pm$1.95	87.59$\pm$4.70	**99.79**$\pm$0.10
Letter	63.38$\pm$17.6	**86.55**$\pm$2.23	60.47$\pm$1.80	70.36$\pm$0.01
Pen-glo	46.86$\pm$1.02	87.25$\pm$1.94	71.93$\pm$1.68	**89.87**$\pm$2.87
Pen-loc	44.11$\pm$6.07	**98.72**$\pm$0.20	64.30$\pm$2.70	90.86$\pm$0.79
Satellite	52.14$\pm$3.08	83.51$\pm$11.9	90.92$\pm$0.16	**94.55**$\pm$0.68
Shuttle	89.37$\pm$5.13	66.29$\pm$1.69	**98.47**$\pm$0.24	78.61$\pm$0.02
Speech	45.61$\pm$3.64	**49.37**$\pm$0.87	47.47$\pm$0.10	46.96$\pm$0.01

Unsupervised outlier detection
real-world benchmarks

ABDA is competitive w.r.t. methods specifically designed for outlier detection

Pattern mining



Single pattern from each leaf L_j^d:

$$\mathcal{P} : \pi_l^d \leq X^d < \pi_h^d$$

for a user-specified threshold θ:

$$\mathcal{S}_{\mathsf{L}_j^d}(\mathcal{P}) \geq \theta$$

Conjunctive patterns

from partitions

$$\mathcal{P}^{\mathsf{N}} = \mathcal{P}_1 \wedge \ldots \wedge \mathcal{P}_{|\mathsf{sc}(\mathsf{N})|}$$

Pattern mining

Ranking by their *relevance* aka ***support***:

$$p_{\mathcal{S}}(\mathcal{P}^i)$$

$\mathcal{P}_1:$ $0.088 \leq \text{Height} < 0.224 \quad \wedge$
$0.158 \leq \text{ShellWeight} < 0.425 \quad (\mathbf{supp}(\mathcal{P}_1) = 0.507)$

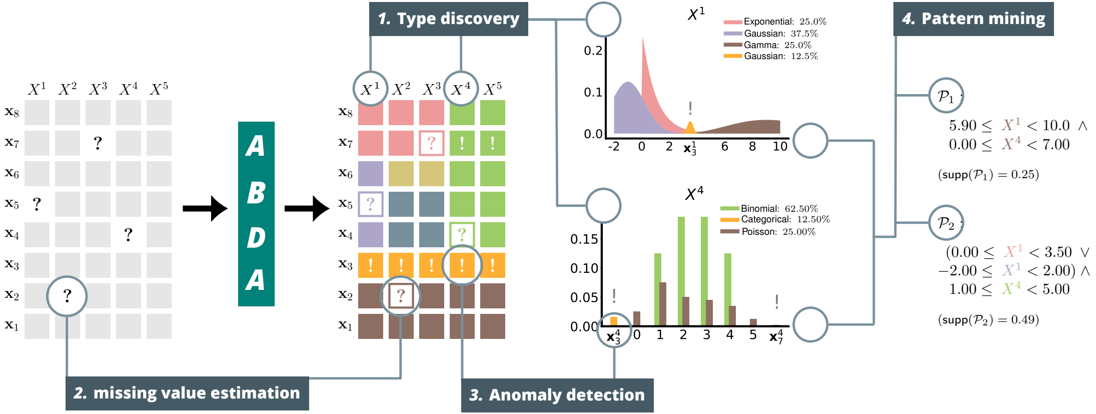
$\mathcal{P}_2:$ $0.483 \leq \text{Length} < 0.684 \quad \wedge$
$0.364 \leq \text{Diameter} < 0.547 \quad (\mathbf{supp}(\mathcal{P}_2) = 0.489)$

$\mathcal{P}_3:$ $0.596 \leq \text{WholeWeight} < 1.040 \quad \wedge\, 0.205 \leq \text{ShuckedWeight} < 0.491 \quad \wedge$
$0.076 \leq \text{VisceraWeight} < 0.281 \quad (\mathbf{supp}(\mathcal{P}_3) = 0.223)$

$\mathcal{P}_4:$ $0.596 \leq \text{WholeWeight} < 1.040 \quad \wedge\, 0.205 \leq \text{ShuckedWeight} < 0.491 \quad \wedge$
$0.076 \leq \text{VisceraWeight} < 0.281 \quad (\mathbf{supp}(\mathcal{P}_4) = 0.202)$

$\mathcal{P}_5:$ $0.313 \leq \text{Length} < 0.554 \quad \wedge \quad 0.230 \leq \text{Diameter} < 0.438 \quad \wedge$
$0.023 \leq \text{Height} < 0.197 \quad \wedge \quad 0.165 \leq \text{WholeWeight} < 0.639 \quad \wedge$
$0.037 \leq \text{ShuckedWeight} < 0.408 \quad \wedge\, 0.019 \leq \text{VisceraWeight} < 0.192 \quad \wedge$
$0.027 \leq \text{ShellWeight} < 0.27 \quad \wedge \quad 5 \leq \text{Rings} < 13 \quad (\mathbf{supp}(\mathcal{P}_5) = 0.111)$

Automatic exploratory data analysis for non-statisticians



for the details visit us at poster #**5839**

code `github.com/probabilistic-learning/abda`

for more on SPNs `github.com/SPFlow/SPFlow`

...and even more `github.com/arranger1044/awesome-spn`