ETH zürich

Significant Pattern Mining for Biomarker Discovery

Felipe Llinares-López

Krupp Symposium: From Machine Learning To Personalized Medicine

MLCB

D-BSSE
Department of Biosystems
Science and Engineering

Machine Learning for Personalized Medicine

Machine Learning and Computational Biology Lab, D-BSSE

Introduction

Significant itemset mining looks for significant multiplicative feature interactions

p features

n samples

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}
$y=0$	1	1	0	0	1	1	0	1	1	1
	1	1	0	1	1	1	0	0	1	0
	1	0	1	0	1	0	0	0	1	1
	1	1	1	0	1	1	1	0	0	1
	0	1	1	0	1	1	0	1	1	0
	1	1	1	0	0	1	0	0	0	1
$y=1$	1	1	1	0	1	1	0	0	1	1
	1	1	1	0	1	1	1	0	1	1
	1	1	1	0	1	1	0	0	1	1
	1	1	1	0	1	1	0	1	0	0
	1	1	1	0	1	1	0	1	1	1
	1	1	0	1	1	1	0	0	1	1

Significant itemset mining looks for significant multiplicative feature interactions

p features

n samples

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}
$y=0$	1	1	0	0	1	1	0	1	1	1
	1	1	0	1	1	1	0	0	1	0
	1	0	1	0	1	0	0	0	1	1
	1	1	1	0	1	1	1	0	0	1
	0	1	1	0	1	1	0	1	1	0
	1	1	1	0	0	1	0	0	0	1
$y=1$	1	1	1	0	1	1	0	0	1	1
	1	1	1	0	1	1	1	0	1	1
	1	1	1	0	1	1	0	0	1	1
	1	1	1	0	1	1	0	1	0	0
	1	1	1	0	1	1	0	1	1	1
	1	1	0	1	1	1	0	0	1	1

$$S = \{1, 3, 5, 6\}$$

Significant itemset mining looks for significant multiplicative feature interactions

p features

n samples

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	$z_{\mathcal{S}}$
$y=0$	1	1	0	0	1	1	0	1	1	1	0
	1	1	0	1	1	1	0	0	1	0	0
	1	0	1	0	1	0	0	0	1	1	0
	1	1	1	0	1	1	1	0	0	1	1
	0	1	1	0	1	1	0	1	1	0	0
	1	1	1	0	0	1	0	0	0	1	0
$y=1$	1	1	1	0	1	1	0	0	1	1	1
	1	1	1	0	1	1	1	0	1	1	1
	1	1	1	0	1	1	0	0	1	1	1
	1	1	1	0	1	1	0	1	0	0	1
	1	1	1	0	1	1	0	1	1	1	1
	1	1	0	1	1	1	0	0	1	1	0

$\mathcal{S} = \{1, 3, 5, 6\}$

$z_{\mathcal{S}} = u_1 u_3 u_5 u_6$

$$z_{\mathcal{S}} = \prod_{j \in \mathcal{S}} u_j$$

$z_{\mathcal{S}}$ is the multiplicative interaction of feature set $\mathcal{S}$

Feature interactions can be enriched in the absence of univariate associations

p features

n samples

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}
$y = 0$	1	1	0	0	1	1	0	1	1	1
	1	1	0	1	1	1	0	0	1	0
	1	0	1	0	1	0	0	0	1	1
	1	1	1	0	1	1	1	0	0	1
	0	1	1	0	1	1	0	1	1	0
	1	1	1	0	0	1	0	0	0	1
$y = 1$	1	1	1	0	1	1	0	0	1	1
	1	1	1	0	1	1	1	0	1	1
	1	1	1	0	1	1	0	0	1	1
	1	1	1	0	1	1	0	1	0	0
	1	1	1	0	1	1	0	1	1	1
	1	1	0	1	1	1	0	0	1	1

Feature interactions can be enriched in the absence of univariate associations

p features; n samples

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}
$y = 0$	1	1	0	0	1	1	0	1	1	1
	1	1	0	1	1	1	0	0	1	0
	1	0	1	0	1	0	0	0	1	1
	1	1	1	0	1	1	1	0	0	1
	0	1	1	0	1	1	0	1	1	0
	1	1	1	0	0	1	0	0	0	1
$y = 1$	1	1	1	0	1	1	0	0	1	1
	1	1	1	0	1	1	1	0	1	1
	1	1	1	0	1	1	0	0	1	1
	1	1	1	0	1	1	0	1	0	0
	1	1	1	0	1	1	0	1	1	1
	1	1	0	1	1	1	0	0	1	1

2×2 contingency table for u_7:

	$u_7 = 1$	$u_7 = 0$	
$y = 1$	1	5	6
$y = 0$	1	5	6
	2	10	12

$\longrightarrow p_{\chi^2}(u_7) = 0.296$

Feature interactions can be enriched in the absence of univariate associations

p features

n samples

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}
$y=0$	1	1	0	0	1	1	0	1	1	1
	1	1	0	1	1	1	0	0	1	0
	1	0	1	0	1	0	0	0	1	1
	1	1	1	0	1	1	1	0	0	1
	0	1	1	0	1	1	0	1	1	0
	1	1	1	0	0	1	0	0	0	1
$y=1$	1	1	1	0	1	1	0	0	1	1
	1	1	1	0	1	1	1	0	1	1
	1	1	1	0	1	1	0	0	1	1
	1	1	1	0	1	1	0	1	0	0
	1	1	1	0	1	1	0	1	1	1
	1	1	0	1	1	1	0	0	1	1

2×2 contingency table for u_7:

	$u_7 = 1$	$u_7 = 0$	
$y = 1$	1	5	6
$y = 0$	1	5	6
	2	10	12

$\longrightarrow p_{\chi^2}(u_7) = 0.296$

Repeat for $u_1, u_2, \ldots, u_{10}$ independently

$$p_{\chi^2}(u_1) = 0.296, \quad p_{\chi^2}(u_6) = 0.296$$
$$p_{\chi^2}(u_2) = 0.296, \quad p_{\chi^2}(u_7) = 0.296$$
$$p_{\chi^2}(u_3) = 0.505, \quad p_{\chi^2}(u_8) = 1.000$$
$$p_{\chi^2}(u_4) = 1.000, \quad p_{\chi^2}(u_9) = 0.505$$
$$p_{\chi^2}(u_5) = 0.296, \quad p_{\chi^2}(u_{10}) = 0.505$$

Feature interactions can be enriched in the absence of univariate associations

p features

n samples

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	$z_{\mathcal{S}_1}$
$y=0$	1	1	0	0	1	1	0	1	1	1	0
	1	1	0	1	1	1	0	0	1	0	0
	1	0	1	0	1	0	0	0	1	1	0
	1	1	1	0	1	1	1	0	0	1	1
	0	1	1	0	1	1	0	1	1	0	0
	1	1	1	0	0	1	0	0	0	1	0
$y=1$	1	1	1	0	1	1	0	0	1	1	1
	1	1	1	0	1	1	1	0	1	1	1
	1	1	1	0	1	1	0	0	1	1	1
	1	1	1	0	1	1	0	1	0	0	1
	1	1	1	0	1	1	0	1	1	1	1
	1	1	0	1	1	1	0	0	1	1	0

2×2 contingency table for $z_{\mathcal{S}_1}$:

	$z_{\mathcal{S}_1}=1$	$z_{\mathcal{S}_1}=0$	
$y=1$	5	1	6
$y=0$	1	5	6
	6	6	12

$\longrightarrow p_{\chi^2}(z_{\mathcal{S}_1}) = 0.021$

Feature interactions can be enriched in the absence of univariate associations

p features

n samples

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	$z_{\mathcal{S}_1}$	$z_{\mathcal{S}_2}$
$y=0$	1	1	0	0	1	1	0	1	1	1	0	1
	1	1	0	1	1	1	0	0	1	0	0	0
	1	0	1	0	1	0	0	0	1	1	0	0
	1	1	1	0	1	1	1	0	0	1	1	0
	0	1	1	0	1	1	0	1	1	0	0	0
	1	1	1	0	0	1	0	0	0	1	0	0
$y=1$	1	1	1	0	1	1	0	0	1	1	1	1
	1	1	1	0	1	1	1	0	1	1	1	1
	1	1	1	0	1	1	0	0	1	1	1	1
	1	1	1	0	1	1	0	1	0	0	1	0
	1	1	1	0	1	1	0	1	1	1	1	1
	1	1	0	1	1	1	0	0	1	1	0	1

2×2 contingency table for $z_{\mathcal{S}_1}$:

	$z_{\mathcal{S}_1} = 1$	$z_{\mathcal{S}_1} = 0$	
$y = 1$	5	1	6
$y = 0$	1	5	6
	6	6	12

$\longrightarrow p_{\chi^2}(z_{\mathcal{S}_1}) = 0.021$

2×2 contingency table for $z_{\mathcal{S}_2}$:

	$z_{\mathcal{S}_2} = 1$	$z_{\mathcal{S}_2} = 0$	
$y = 1$	5	1	6
$y = 0$	1	5	6
	6	6	12

$\longrightarrow p_{\chi^2}(z_{\mathcal{S}_2}) = 0.021$

Significant itemset mining has many applications in personalized medicine

p features

n samples

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	z_S
$y=0$	1	1	0	0	1	1	0	1	1	1	0
	1	1	0	1	1	1	0	0	1	0	0
	1	0	1	0	1	0	0	0	1	1	0
	1	1	1	0	1	1	1	0	0	1	1
	0	1	1	0	1	1	0	1	1	0	0
	1	1	1	0	0	1	0	0	0	1	0
$y=1$	1	1	1	0	1	1	0	0	1	1	1
	1	1	1	0	1	1	1	0	1	1	1
	1	1	1	0	1	1	0	0	1	1	1
	1	1	1	0	1	1	0	1	0	0	1
	1	1	1	0	1	1	0	1	1	1	1
	1	1	0	1	1	1	0	0	1	1	0

Significant itemset mining has many applications in personalized medicine

p genomic markers

n individuals

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	z_S
n_2 controls	1	1	0	0	1	1	0	1	1	1	0
n_2 controls	1	1	0	1	1	1	0	0	1	0	0
n_2 controls	1	0	1	0	1	0	0	0	1	1	0
n_2 controls	1	1	1	0	1	1	1	0	0	1	1
n_2 controls	0	1	1	0	1	1	0	1	1	0	0
n_2 controls	1	1	1	0	0	1	0	0	0	1	0
n_1 cases	1	1	1	0	1	1	0	0	1	1	1
n_1 cases	1	1	1	0	1	1	1	0	1	1	1
n_1 cases	1	1	1	0	1	1	0	0	1	1	1
n_1 cases	1	1	1	0	1	1	0	1	0	0	1
n_1 cases	1	1	1	0	1	1	0	1	1	1	1
n_1 cases	1	1	0	1	1	1	0	0	1	1	0

- Association studies in genetics
  - SNPs
  - (Discretized) gene expression
  - Epigenetics (e.g methylation)

Significant itemset mining has many applications in personalized medicine

p TF binding motifs

n genes		y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	z_S
	n_2 not up-regulated genes	→|	1	1	0	0	1	1	0	1	1	1	0
		→|	1	1	0	1	1	1	0	0	1	0	0
		→|	1	0	1	0	1	0	0	0	1	1	0
		→|	1	1	1	0	1	1	1	0	0	1	1
		→|	0	1	1	0	1	1	0	1	1	0	0
		→|	1	1	1	0	0	1	0	0	0	1	0
	n_1 up-regulated genes	↑	1	1	1	0	1	1	0	0	1	1	1
		↑	1	1	1	0	1	1	1	0	1	1	1
		↑	1	1	1	0	1	1	0	0	1	1	1
		↑	1	1	1	0	1	1	0	1	0	0	1
		↑	1	1	1	0	1	1	0	1	1	1	1
		↑	1	1	0	1	1	1	0	0	1	1	0

- Association studies in genetics
  - SNPs
  - (Discretized) gene expression
  - Epigenetics (e.g methylation)
- Functional genomics
  - Combinational transcription factor (TF) binding

Significant itemset mining has many applications in personalized medicine

p chromatin marks

n genomic regions (e.g. 200 bp long)

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	z_S
n_2 non-enhancer regions	1	1	0	0	1	1	0	1	1	1	0
	1	1	0	1	1	1	0	0	1	0	0
	1	0	1	0	1	0	0	0	1	1	0
	1	1	1	0	1	1	1	0	0	1	1
	0	1	1	0	1	1	0	1	1	0	0
	1	1	1	0	0	1	0	0	0	1	0
n_1 enhancer regions	1	1	1	0	1	1	0	0	1	1	1
	1	1	1	0	1	1	1	0	1	1	1
	1	1	1	0	1	1	0	0	1	1	1
	1	1	1	0	1	1	0	1	0	0	1
	1	1	1	0	1	1	0	1	1	1	1
	1	1	0	1	1	1	0	0	1	1	0

- Association studies in genetics
  - SNPs
  - (Discretized) gene expression
  - Epigenetics (e.g methylation)
- Functional genomics
  - Combinational transcription factor (TF) binding
  - Mapping chromatin marks to genomic function

Significant itemset mining has many applications in personalized medicine

p symptoms

n patients

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	z_S
good prognosis	1	1	0	0	1	1	0	1	1	1	0
good prognosis	1	1	0	1	1	1	0	0	1	0	0
good prognosis	1	0	1	0	1	0	0	0	1	1	0
good prognosis	1	1	1	0	1	1	1	0	0	1	1
good prognosis	0	1	1	0	1	1	0	1	1	0	0
good prognosis	1	1	1	0	0	1	0	0	0	1	0
bad prognosis	1	1	1	0	1	1	0	0	1	1	1
bad prognosis	1	1	1	0	1	1	1	0	1	1	1
bad prognosis	1	1	1	0	1	1	0	0	1	1	1
bad prognosis	1	1	1	0	1	1	0	1	0	0	1
bad prognosis	1	1	1	0	1	1	0	1	1	1	1
bad prognosis	1	1	0	1	1	1	0	0	1	1	0

- Association studies in genetics
  - SNPs
  - (Discretized) gene expression
  - Epigenetics (e.g methylation)
- Functional genomics
  - Combinational transcription factor (TF) binding
  - Mapping chromatin marks to genomic function
- Mining clinical databases

Significant itemset mining poses both computational and statistical challenges

p features

n samples

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	z_S
	1	1	0	0	1	1	0	1	1	1	0
	1	1	0	1	1	1	0	0	1	0	0
	1	0	1	0	1	0	0	0	1	1	0
	1	1	1	0	1	1	1	0	0	1	1
	0	1	1	0	1	1	0	1	1	0	0
	1	1	1	0	0	1	0	0	0	1	0
	1	1	1	0	1	1	0	0	1	1	1
	1	1	1	0	1	1	1	0	1	1	1
	1	1	1	0	1	1	0	0	1	1	1
	1	1	1	0	1	1	0	1	0	0	1
	1	1	1	0	1	1	0	1	1	1	1
	1	1	0	1	1	1	0	0	1	1	0

Significant itemset mining poses both computational and statistical challenges

p features

n samples

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	$z_{\mathcal{S}}$
	1	1	0	0	1	1	0	1	1	1	0
	1	1	0	1	1	1	0	0	1	0	0
	1	0	1	0	1	0	0	0	1	1	0
	1	1	1	0	1	1	1	0	0	1	1
	0	1	1	0	1	1	0	1	1	0	0
	1	1	1	0	0	1	0	0	0	1	0
	1	1	1	0	1	1	0	0	1	1	1
	1	1	1	0	1	1	1	0	1	1	1
	1	1	1	0	1	1	0	0	1	1	1
	1	1	1	0	1	1	0	1	0	0	1
	1	1	1	0	1	1	0	1	1	1	1
	1	1	0	1	1	1	0	0	1	1	0

In a dataset with p binary features there are up to 2^p feature interactions $\mathcal{S} \subseteq \{1,2,\ldots,p\}$

Significant itemset mining poses both computational and statistical challenges

p features

n samples

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	$z_{\mathcal{S}}$
	1	1	0	0	1	1	0	1	1	1	0
	1	1	0	1	1	1	0	0	1	0	0
	1	0	1	0	1	0	0	0	1	1	0
	1	1	1	0	1	1	1	0	0	1	1
	0	1	1	0	1	1	0	1	1	0	0
	1	1	1	0	0	1	0	0	0	1	0
	1	1	1	0	1	1	0	0	1	1	1
	1	1	1	0	1	1	1	0	1	1	1
	1	1	1	0	1	1	0	0	1	1	1
	1	1	1	0	1	1	0	1	0	0	1
	1	1	1	0	1	1	0	1	1	1	1
	1	1	0	1	1	1	0	0	1	1	0

In a dataset with p binary features there are up to 2^p feature interactions $\mathcal{S} \subseteq \{1,2,\dots,p\}$

- For comparison (see [1, Appendix C.4]):
  - $p = 266$: # of feature interactions ≈ # of electrons in the observable universe

Significant itemset mining poses both computational and statistical challenges

p features

n samples

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	$z_{\mathcal{S}}$
	1	1	0	0	1	1	0	1	1	1	0
	1	1	0	1	1	1	0	0	1	0	0
	1	0	1	0	1	0	0	0	1	1	0
	1	1	1	0	1	1	1	0	0	1	1
	0	1	1	0	1	1	0	1	1	0	0
	1	1	1	0	0	1	0	0	0	1	0
	1	1	1	0	1	1	0	0	1	1	1
	1	1	1	0	1	1	1	0	1	1	1
	1	1	1	0	1	1	0	0	1	1	1
	1	1	1	0	1	1	0	1	0	0	1
	1	1	1	0	1	1	0	1	1	1	1
	1	1	0	1	1	1	0	0	1	1	0

In a dataset with p binary features there are up to 2^p feature interactions $\mathcal{S} \subseteq \{1,2,\ldots,p\}$

- For comparison (see [1, Appendix C.4]):
  - $p = 266$: # of feature interactions ≈ # of electrons in the observable universe
- This leads to two fundamental challenges:
  - Computational
  - Statistical (*multiple comparisons problem*)

There exist *untestable* feature interactions that cannot cause false positives

p features

n samples

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	z_S
	1	1	0	0	1	1	0	1	1	1	0
	1	1	0	1	1	1	0	0	1	0	0
	1	0	1	0	1	0	0	0	1	1	0
	1	1	1	0	1	1	1	0	0	1	1
	0	1	1	0	1	1	0	1	1	0	0
	1	1	1	0	0	1	0	0	0	1	0
	1	1	1	0	1	1	0	0	1	1	1
	1	1	1	0	1	1	1	0	1	1	1
	1	1	1	0	1	1	0	0	1	1	1
	1	1	1	0	1	1	0	1	0	0	1
	1	1	1	0	1	1	0	1	1	1	1
	1	1	0	1	1	1	0	0	1	1	0

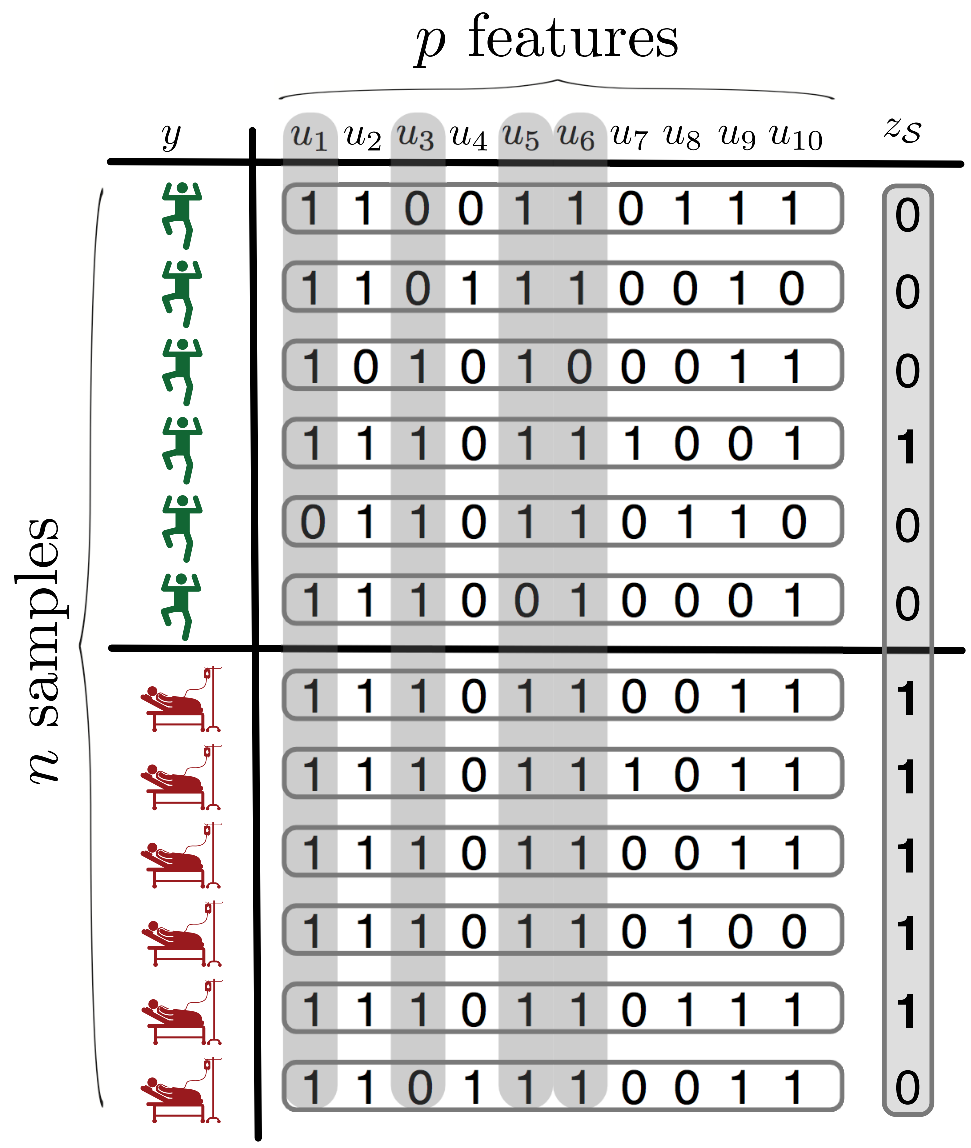
There exist *untestable* feature interactions that cannot cause false positives

p features; n samples

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	z_S
	1	1	0	0	1	1	0	1	1	1	0
	1	1	0	1	1	1	0	0	1	0	0
	1	0	1	0	1	0	0	0	1	1	0
	1	1	1	0	1	1	1	0	0	1	1
	0	1	1	0	1	1	0	1	1	0	0
	1	1	1	0	0	1	0	0	0	1	0
	1	1	1	0	1	1	0	0	1	1	1
	1	1	1	0	1	1	1	0	1	1	1
	1	1	1	0	1	1	0	0	1	1	1
	1	1	1	0	1	1	0	1	0	0	1
	1	1	1	0	1	1	0	1	1	1	1
	1	1	0	1	1	1	0	0	1	1	0

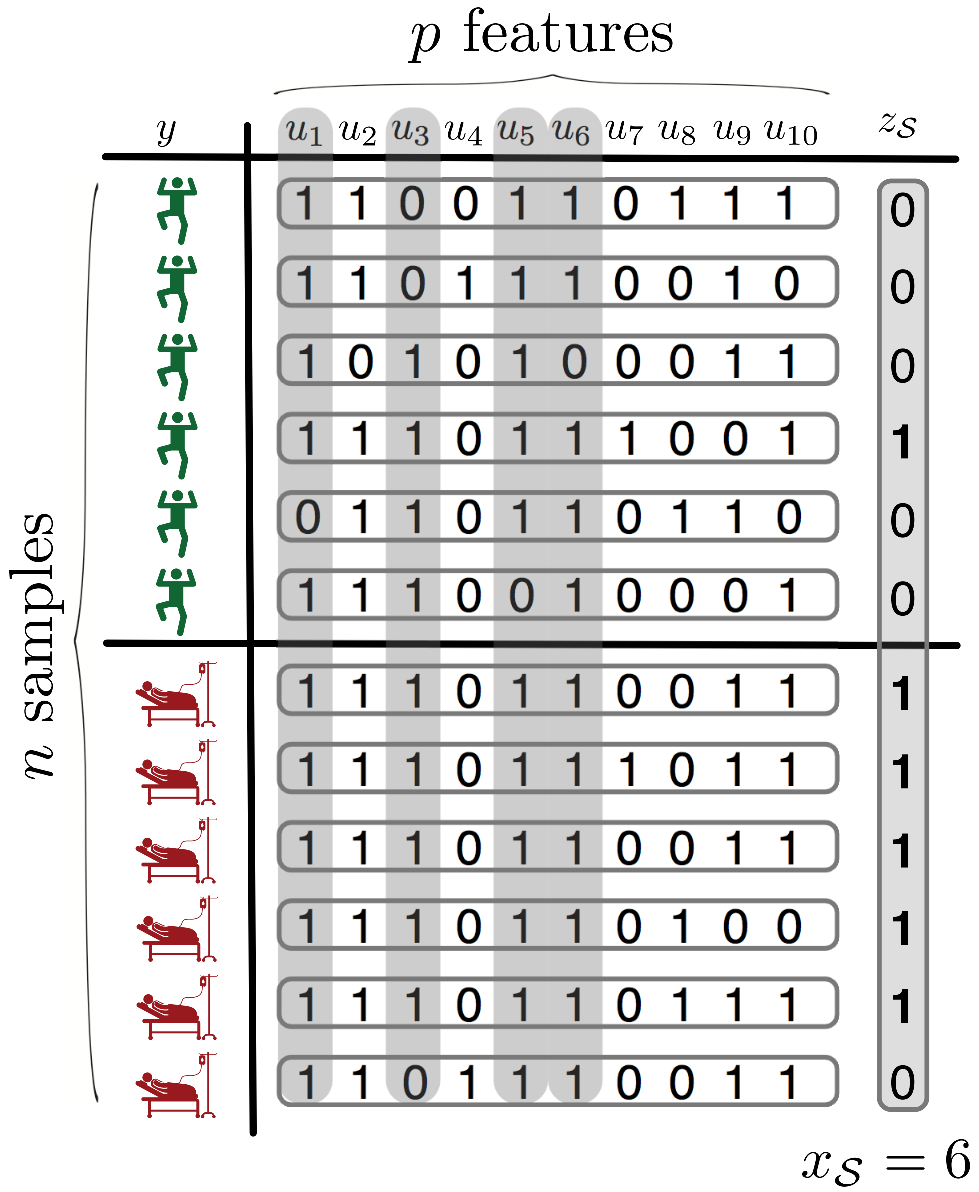
- For discrete test statistics, p-values cannot be arbitrarily small

There exist *untestable* feature interactions that cannot cause false positives



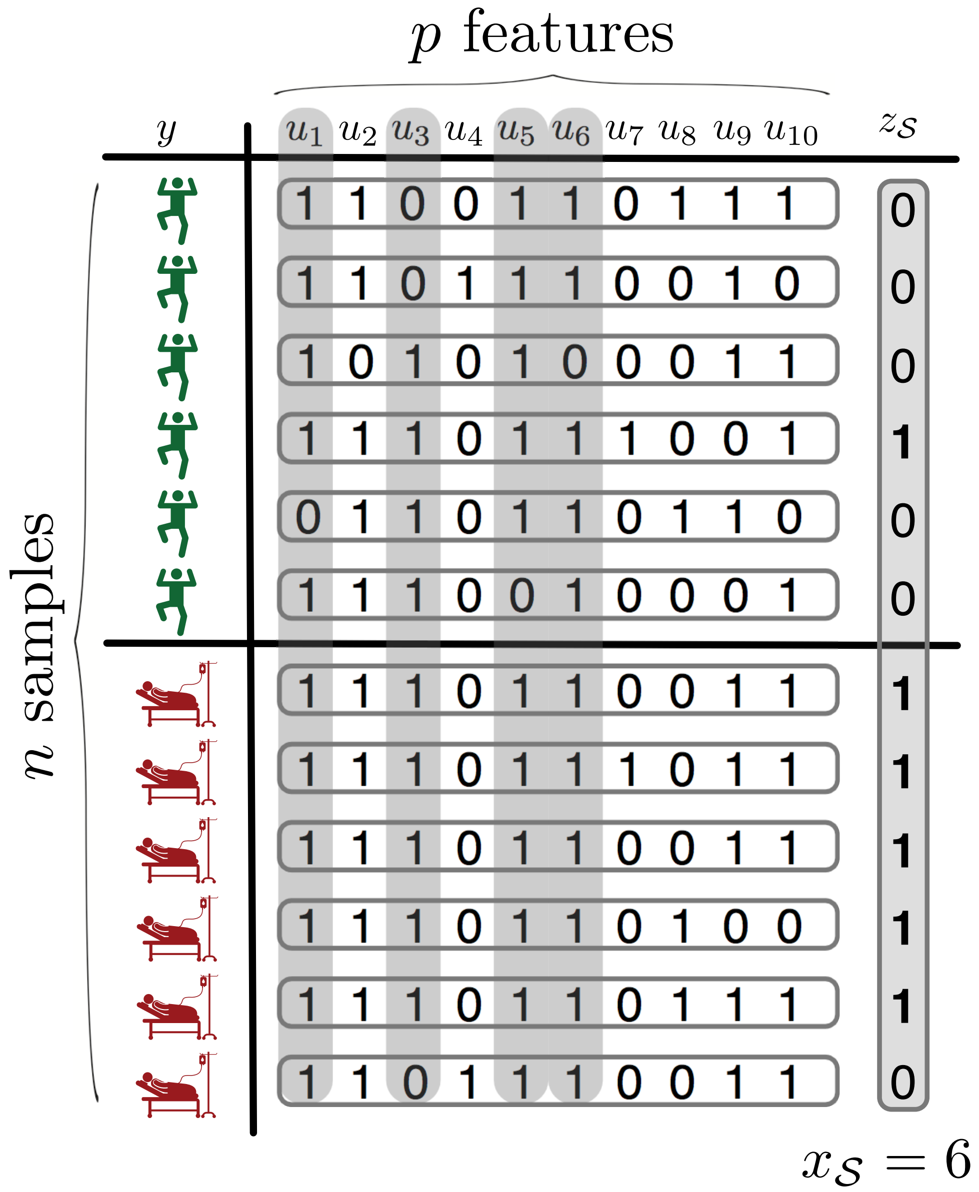
- For discrete test statistics, p-values cannot be arbitrarily small
- A minimum attainable p-value can be computed as a function of the margins of the contingency table

There exist *untestable* feature interactions that cannot cause false positives



- For discrete test statistics, p-values cannot be arbitrarily small
- A minimum attainable p-value can be computed as a function of the margins of the contingency table
- For each $\mathcal{S} \subseteq \{1,2,\ldots,p\}$, its minimum attainable p-value $\Psi(x_{\mathcal{S}})$ is a function of the support $x_{\mathcal{S}}$ of the interaction
  - $x_{\mathcal{S}} \equiv$ # of samples with $z_{\mathcal{S}} = 1$

There exist *untestable* feature interactions that cannot cause false positives



- For discrete test statistics, p-values cannot be arbitrarily small
- A minimum attainable p-value can be computed as a function of the margins of the contingency table
- For each $\mathcal{S} \subseteq \{1,2,\dots,p\}$, its minimum attainable p-value $\Psi(x_{\mathcal{S}})$ is a function of the support $x_{\mathcal{S}}$ of the interaction
  - $x_{\mathcal{S}} \equiv$ # of samples with $z_{\mathcal{S}} = 1$
- **[Tarone, Biometrics 1990]** *Untestable* feature interactions $\mathcal{S}$ for which $\Psi(x_{\mathcal{S}}) > \delta$ can neither be significant nor cause a false positive
  - $\delta \equiv$ significance threshold

LAMP (Terada et al., PNAS 2013) tackles both challenges using *testability*

$\emptyset$

u_1 u_2 u_3 u_4 u_5

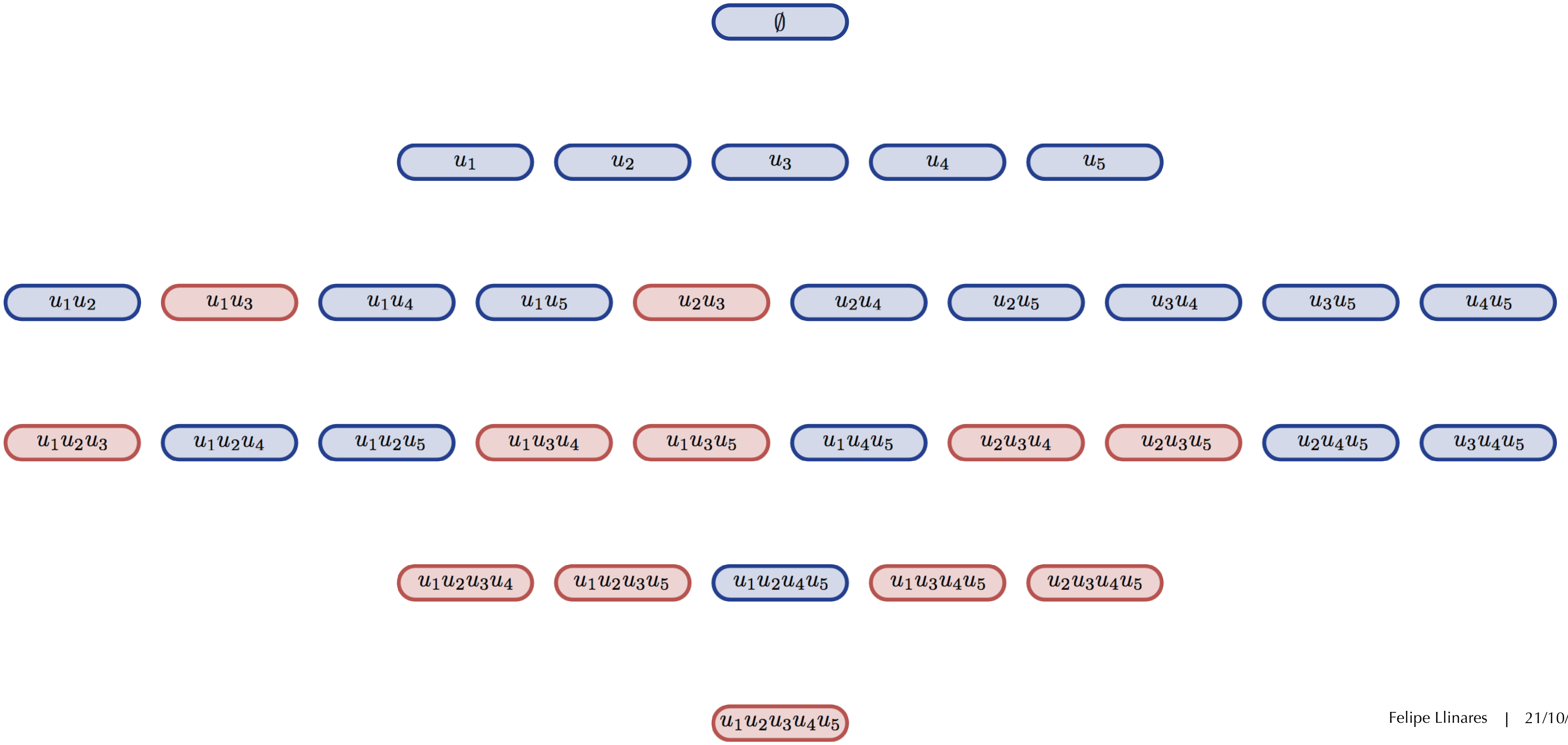
u_1u_2 u_1u_3 u_1u_4 u_1u_5 u_2u_3 u_2u_4 u_2u_5 u_3u_4 u_3u_5 u_4u_5

$u_1u_2u_3$ $u_1u_2u_4$ $u_1u_2u_5$ $u_1u_3u_4$ $u_1u_3u_5$ $u_1u_4u_5$ $u_2u_3u_4$ $u_2u_3u_5$ $u_2u_4u_5$ $u_3u_4u_5$

$u_1u_2u_3u_4$ $u_1u_2u_3u_5$ $u_1u_2u_4u_5$ $u_1u_3u_4u_5$ $u_2u_3u_4u_5$

$u_1u_2u_3u_4u_5$

LAMP (Terada et al., PNAS 2013) tackles both challenges using *testability*



LAMP (Terada et al., PNAS 2013) tackles both challenges using *testability*

$\emptyset$

u_1 u_2 u_3 u_4 u_5

u_1u_2 u_1u_3 u_1u_4 u_1u_5 u_2u_3 u_2u_4 u_2u_5 u_3u_4 u_3u_5 u_4u_5

$u_1u_2u_3$ $u_1u_2u_4$ $u_1u_2u_5$ $u_1u_3u_4$ $u_1u_3u_5$ $u_1u_4u_5$ $u_2u_3u_4$ $u_2u_3u_5$ $u_2u_4u_5$ $u_3u_4u_5$

$u_1u_2u_3u_4$ $u_1u_2u_3u_5$ $u_1u_2u_4u_5$ $u_1u_3u_4u_5$ $u_2u_3u_4u_5$

$u_1u_2u_3u_4u_5$

LAMP (Terada et al., PNAS 2013) tackles both challenges using *testability*

1) $x_{\mathcal{S}} \geq x_{\mathcal{S}'} \Rightarrow \Psi(x_{\mathcal{S}}) \leq \Psi(x_{\mathcal{S}'})$

$\emptyset$

u_1 u_2 u_3 u_4 u_5

u_1u_2 u_1u_3 u_1u_4 u_1u_5 u_2u_3 u_2u_4 u_2u_5 u_3u_4 u_3u_5 u_4u_5

$u_1u_2u_3$ $u_1u_2u_4$ $u_1u_2u_5$ $u_1u_3u_4$ $u_1u_3u_5$ $u_1u_4u_5$ $u_2u_3u_4$ $u_2u_3u_5$ $u_2u_4u_5$ $u_3u_4u_5$

$u_1u_2u_3u_4$ $u_1u_2u_3u_5$ $u_1u_2u_4u_5$ $u_1u_3u_4u_5$ $u_2u_3u_4u_5$

$u_1u_2u_3u_4u_5$

LAMP (Terada et al., PNAS 2013) tackles both challenges using *testability*

1) $x_{\mathcal{S}} \geq x_{\mathcal{S}'} \Rightarrow \Psi(x_{\mathcal{S}}) \leq \Psi(x_{\mathcal{S}'})$

2) $\mathcal{S} \subseteq \mathcal{S}' \Rightarrow x_{\mathcal{S}} \geq x_{\mathcal{S}'}$

$\emptyset$

u_1 u_2 u_3 u_4 u_5

u_1u_2 u_1u_3 u_1u_4 u_1u_5 u_2u_3 u_2u_4 u_2u_5 u_3u_4 u_3u_5 u_4u_5

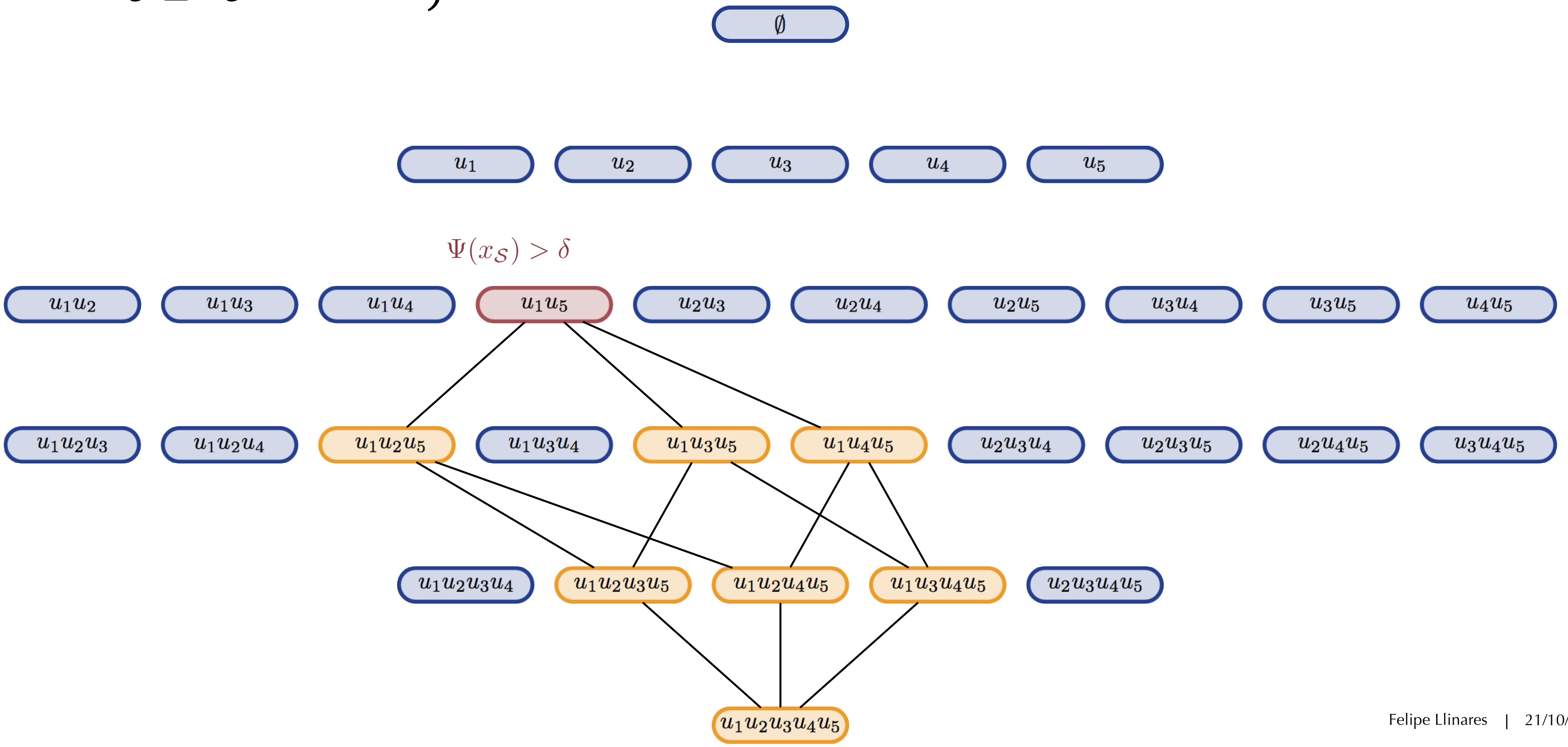
$u_1u_2u_3$ $u_1u_2u_4$ $u_1u_2u_5$ $u_1u_3u_4$ $u_1u_3u_5$ $u_1u_4u_5$ $u_2u_3u_4$ $u_2u_3u_5$ $u_2u_4u_5$ $u_3u_4u_5$

$u_1u_2u_3u_4$ $u_1u_2u_3u_5$ $u_1u_2u_4u_5$ $u_1u_3u_4u_5$ $u_2u_3u_4u_5$

$u_1u_2u_3u_4u_5$

LAMP (Terada et al., PNAS 2013) tackles both challenges using *testability*

$$\left.\begin{array}{l} 1)\ x_{\mathcal{S}} \geq x_{\mathcal{S}'} \Rightarrow \Psi(x_{\mathcal{S}}) \leq \Psi(x_{\mathcal{S}'}) \\ 2)\ \mathcal{S} \subseteq \mathcal{S}' \Rightarrow x_{\mathcal{S}} \geq x_{\mathcal{S}'} \end{array}\right\} \quad \text{If } \mathcal{S} \subseteq \mathcal{S}', \Psi(x_{\mathcal{S}}) > \delta \Rightarrow \Psi(x_{\mathcal{S}'}) > \delta$$



Accounting for the dependence between feature interactions

Exponentially-many combinatorial feature interactions are statistically dependent

p features

n samples

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	z_S
	1	1	0	0	1	1	0	1	1	1	1
	1	1	0	1	1	1	0	0	1	0	0
	1	0	1	0	1	0	0	0	1	1	0
	1	1	1	0	1	1	1	0	0	1	0
	0	1	1	0	1	1	0	1	1	0	0
	1	1	1	0	0	1	0	0	0	1	0
	1	1	1	0	1	1	0	0	1	1	1
	1	1	1	0	1	1	1	0	1	1	1
	1	1	1	0	1	1	0	0	1	1	1
	1	1	1	0	1	1	0	1	0	0	0
	1	1	1	0	1	1	0	1	1	1	1
	1	1	0	1	1	1	0	0	1	1	1

$S = \{2, 9, 10\}$

Exponentially-many combinatorial feature interactions are statistically dependent

p features

n samples

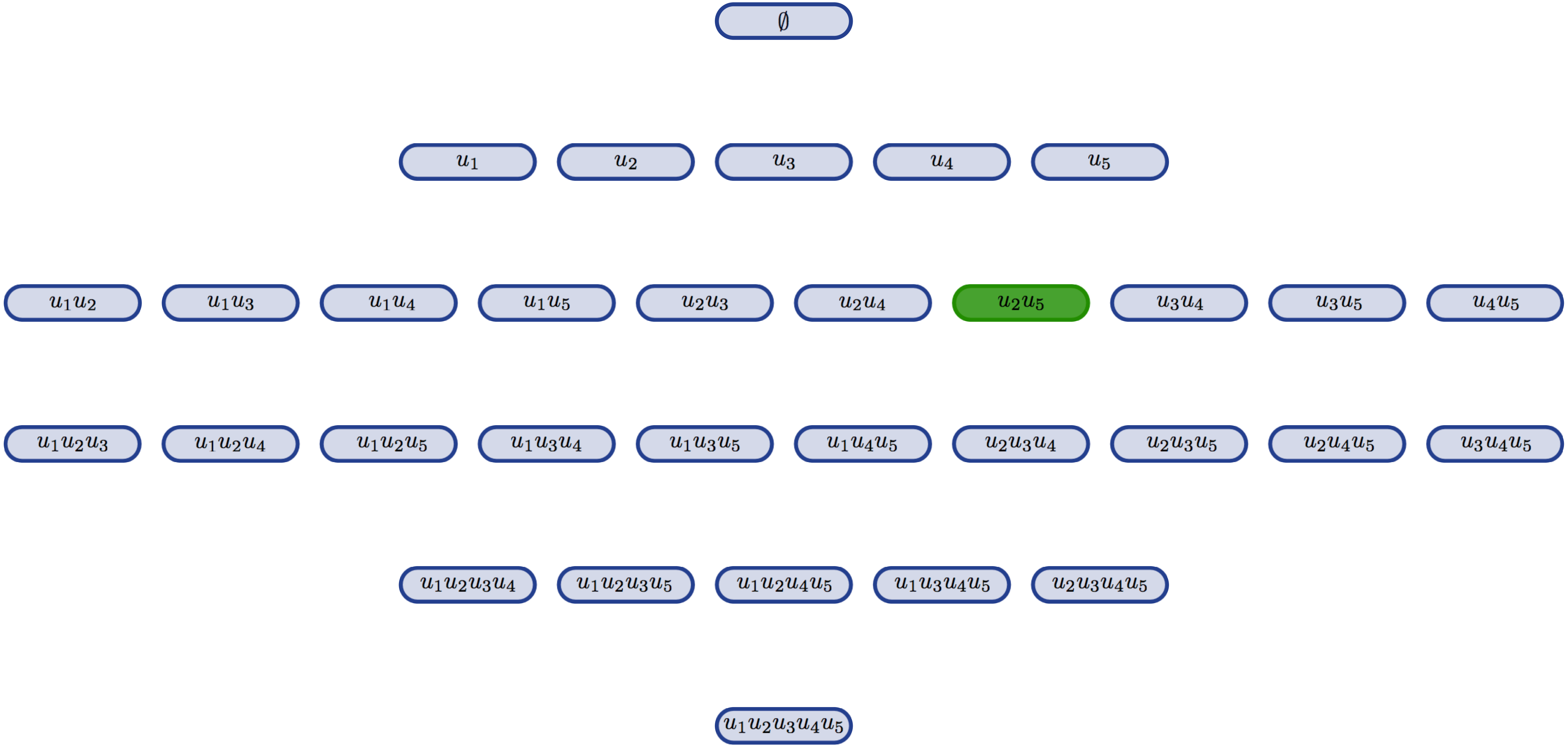
y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	$z_{\mathcal{S}}$	$z_{\mathcal{S}'}$
	1	1	0	0	1	1	0	1	1	1	1	?
	1	1	0	1	1	1	0	0	1	0	0	0
	1	0	1	0	1	0	0	0	1	1	0	0
	1	1	1	0	1	1	1	0	0	1	0	0
	0	1	1	0	1	1	0	1	1	0	0	0
	1	1	1	0	0	1	0	0	0	1	0	0
	1	1	1	0	1	1	0	0	1	1	1	?
	1	1	1	0	1	1	1	0	1	1	1	?
	1	1	1	0	1	1	0	0	1	1	1	?
	1	1	1	0	1	1	0	1	0	0	0	0
	1	1	1	0	1	1	0	1	1	1	1	?
	1	1	0	1	1	1	0	0	1	1	1	?

$\mathcal{S} = \{2, 9, 10\} \rightarrow \mathcal{S}' = \mathcal{S} \cup \{7\}$

$z_{\mathcal{S}'} = u_7 z_{\mathcal{S}}$

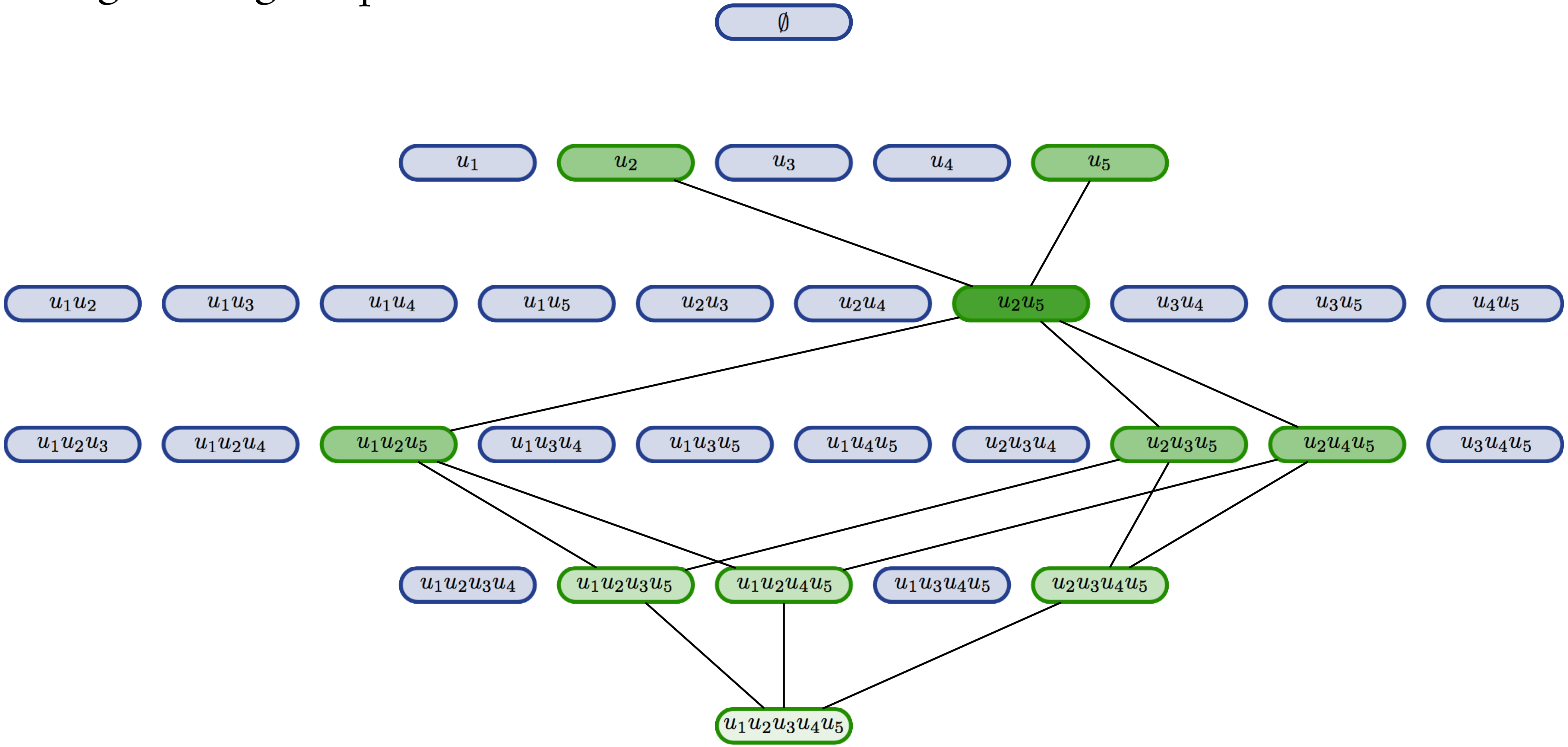
Superset/subset relationships induce statistical dependence between feature interactions

Exponentially-many combinatorial feature interactions are statistically dependent



Exponentially-many combinatorial feature interactions are statistically dependent

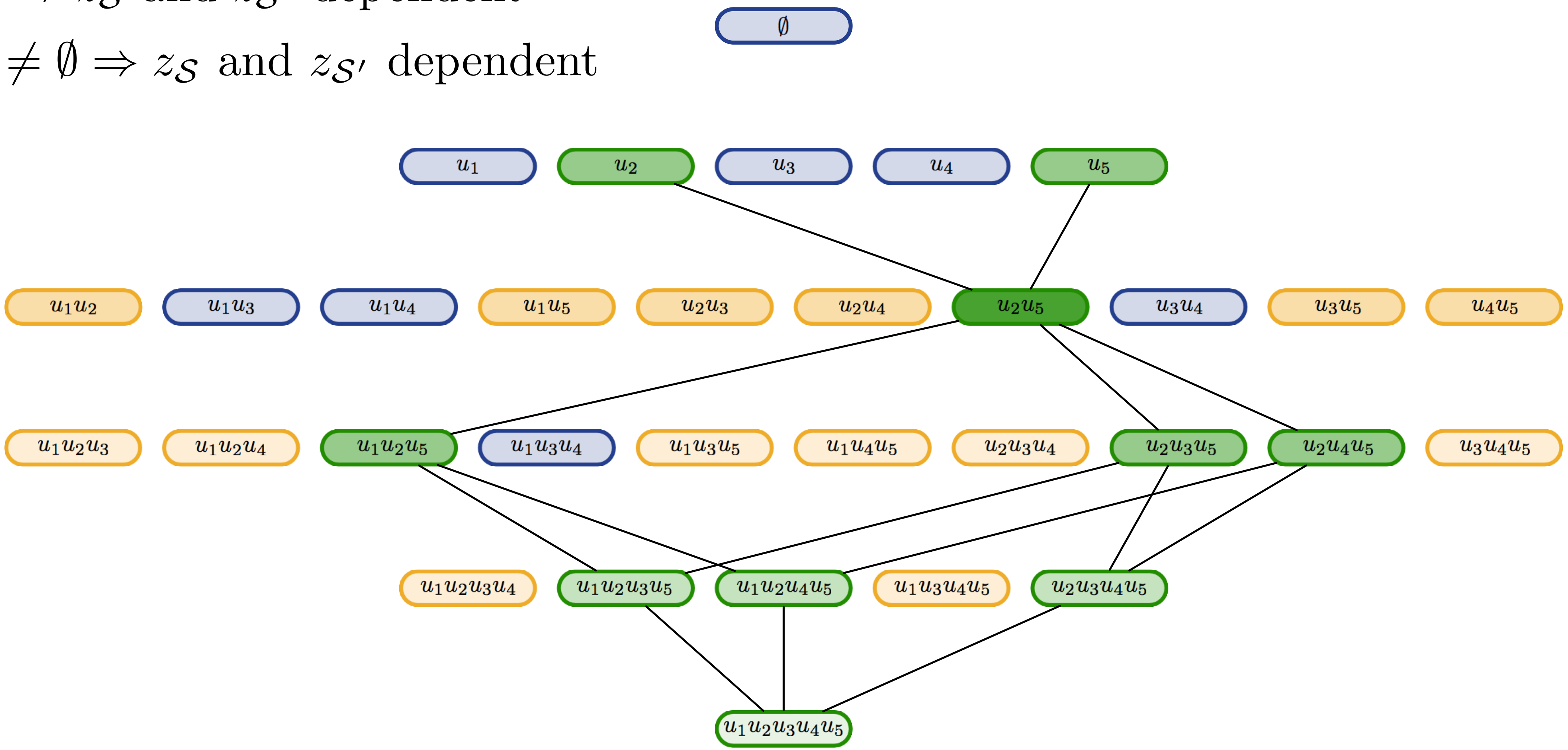
$\mathcal{S} \subseteq \mathcal{S}' \Rightarrow z_{\mathcal{S}}$ and $z_{\mathcal{S}'}$ dependent



Exponentially-many combinatorial feature interactions are statistically dependent

$\mathcal{S} \subseteq \mathcal{S}' \Rightarrow z_{\mathcal{S}}$ and $z_{\mathcal{S}'}$ dependent

$\mathcal{S} \cap \mathcal{S}' \neq \emptyset \Rightarrow z_{\mathcal{S}}$ and $z_{\mathcal{S}'}$ dependent



Ignoring the dependence between candidate interactions leads to a loss of power

Ignoring the dependence between candidate interactions leads to a loss of power

- We deem a combinatorial feature interaction $\mathcal{S}$ significant if $p(z_{\mathcal{S}}) \leq \delta$

Ignoring the dependence between candidate interactions leads to a loss of power

- We deem a combinatorial feature interaction $\mathcal{S}$ significant if $p(z_{\mathcal{S}}) \leq \delta$
- **FWER control at level α:** $\delta^* = \max \delta \;\; \text{s.t.} \;\; FWER(\delta) = \Pr(FP(\delta) \neq 0) \leq \alpha$

Ignoring the dependence between candidate interactions leads to a loss of power

- We deem a combinatorial feature interaction $\mathcal{S}$ significant if $p(z_{\mathcal{S}}) \leq \delta$
- **FWER control at level α:** $\delta^* = \max \delta \ \text{ s.t. } \ FWER(\delta) = \Pr(FP(\delta) \neq 0) \leq \alpha$

$FWER(\delta)$ unknown!

Ignoring the dependence between candidate interactions leads to a loss of power

- We deem a combinatorial feature interaction $\mathcal{S}$ significant if $p(z_{\mathcal{S}}) \leq \delta$

- **FWER control at level α:** $\delta^* = \max \delta \ \text{ s.t. } \ FWER(\delta) = \Pr(FP(\delta) \neq 0) \leq \alpha$

$FWER(\delta)$ unknown!

- Bonferroni correction: $FWER_{bon}(\delta) = 2^p \delta$
  - Assumes all feature interactions can be significant
  - Assumes all feature interactions are statistically independent

Ignoring the dependence between candidate interactions leads to a loss of power

- We deem a combinatorial feature interaction $\mathcal{S}$ significant if $p(z_{\mathcal{S}}) \leq \delta$

- **FWER control at level $\boldsymbol{\alpha}$:** $\delta^* = \max \delta \;\; \text{s.t.} \;\; FWER(\delta) = \Pr(FP(\delta) \neq 0) \leq \alpha$

$FWER(\delta)$ unknown!

- Bonferroni correction: $FWER_{bon}(\delta) = 2^p \delta$
  - Assumes all feature interactions can be significant
  - Assumes all feature interactions are statistically independent

- Testability-based Bonferroni correction: $FWER_{tar}(\delta) = m(\delta)\delta$
  - Only considers *testable* feature interactions, $m(\delta) \ll 2^p$
  - Assumes all feature interactions are statistically independent

Ignoring the dependence between candidate interactions leads to a loss of power

- We deem a combinatorial feature interaction $\mathcal{S}$ significant if $p(z_{\mathcal{S}}) \leq \delta$
- **FWER control at level α:** $\delta^* = \max \delta \;\; \text{s.t.} \;\; FWER(\delta) = \Pr(FP(\delta) \neq 0) \leq \alpha$

$FWER(\delta)$ unknown!

- Bonferroni correction: $FWER_{bon}(\delta) = 2^p \delta$
  - Assumes all feature interactions can be significant
  - Assumes all feature interactions are statistically independent
- Testability-based Bonferroni correction: $FWER_{tar}(\delta) = m(\delta)\delta$
  - Only considers *testable* feature interactions, $m(\delta) \ll 2^p$
  - Assumes all feature interactions are statistically independent

Loss of power: $\delta^*_{bon} \ll \delta^*_{tar} < \delta^*$

Fast and Memory-Efficient Significant Pattern Mining via Permutation Testing

Felipe Llinares-López
D-BSSE, ETH Zürich
felipe.llinares@bsse.ethz.ch

Mahito Sugiyama
ISIR, Osaka University
JST, PRESTO
mahito@ar.sanken.osaka-u.ac.jp

Laetitia Papaxanthos
D-BSSE, ETH Zürich
laetitia.papaxanthos@bsse.ethz.ch

Karsten M. Borgwardt
D-BSSE, ETH Zürich
karsten.borgwardt@bsse.ethz.ch

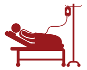
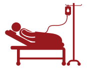
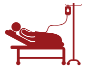
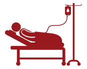
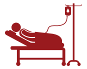
Published at KDD 2015

Goal: Develop a new significant pattern mining algorithm that takes dependence into feature interactions into account

Permutation-testing can be used for accurate estimation of the FWER

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}
	1	1	0	0	1	1	0	1	1	1
	1	1	0	1	1	1	0	0	1	0
	1	0	1	0	1	0	0	0	1	1
	1	1	1	0	1	1	1	0	0	1
	0	1	1	0	1	1	0	1	1	0
	1	1	1	0	0	1	0	0	0	1
	1	1	1	0	1	1	0	0	1	1
	1	1	1	0	1	1	1	0	1	1
	1	1	1	0	1	1	0	0	1	1
	1	1	1	0	1	1	0	1	0	0
	1	1	1	0	1	1	0	1	1	1
	1	1	0	1	1	1	0	0	1	1

Permutation-testing can be used for accurate estimation of the FWER

$\tilde{y}$	$\tilde{u}_1$	$\tilde{u}_2$	$\tilde{u}_3$	$\tilde{u}_4$	$\tilde{u}_5$	$\tilde{u}_6$	$\tilde{u}_7$	$\tilde{u}_8$	$\tilde{u}_9$	$\tilde{u}_{10}$
	1	1	0	0	1	1	0	1	1	1
	1	1	0	1	1	1	0	0	1	0
	1	0	1	0	1	0	0	0	1	1
	1	1	1	0	1	1	1	0	0	1
	0	1	1	0	1	1	0	1	1	0
	1	1	1	0	0	1	0	0	0	1
	1	1	1	0	1	1	0	0	1	1
	1	1	1	0	1	1	1	0	1	1
	1	1	1	0	1	1	0	0	1	1
	1	1	1	0	1	1	0	1	0	0
	1	1	1	0	1	1	0	1	1	1
	1	1	0	1	1	1	0	0	1	1

$$(\tilde{\mathbf{u}}, \tilde{y}) = (\mathbf{u}, \texttt{randperm}(y))$$

$$(\tilde{\mathbf{u}}, \tilde{y}) \sim p(\mathbf{u})p(y) \Leftrightarrow \tilde{\mathbf{u}} \text{ and } \tilde{y} \text{ independent}$$

ETH zürich

Permutation-testing can be used for accurate estimation of the FWER

$\tilde{y}$	$\tilde{u}_1$	$\tilde{u}_2$	$\tilde{u}_3$	$\tilde{u}_4$	$\tilde{u}_5$	$\tilde{u}_6$	$\tilde{u}_7$	$\tilde{u}_8$	$\tilde{u}_9$	$\tilde{u}_{10}$
	1	1	0	0	1	1	0	1	1	1
	1	1	0	1	1	1	0	0	1	0
	1	0	1	0	1	0	0	0	1	1
	1	1	1	0	1	1	1	0	0	1
	0	1	1	0	1	1	0	1	1	0
	1	1	1	0	0	1	0	0	0	1
	1	1	1	0	1	1	0	0	1	1
	1	1	1	0	1	1	1	0	1	1
	1	1	1	0	1	1	0	0	1	1
	1	1	1	0	1	1	0	1	0	0
	1	1	1	0	1	1	0	1	1	1
	1	1	0	1	1	1	0	0	1	1

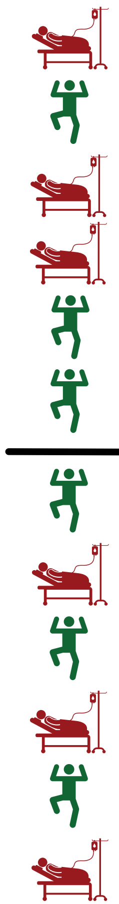
2^p feature interactions

	$z_{\mathcal{S}_1}$	$z_{\mathcal{S}_2}$	$z_{\mathcal{S}_3}$	$\cdots$	$z_{\mathcal{S}_{2^p-1}}$	$z_{\mathcal{S}_{2^p}}$
1	$p^{(1)}(z_{\mathcal{S}_1})$	$p^{(1)}(z_{\mathcal{S}_2})$	$p^{(1)}(z_{\mathcal{S}_3})$	$\cdots$	$p^{(1)}(z_{\mathcal{S}_{2^p-1}})$	$p^{(1)}(z_{\mathcal{S}_{2^p}})$

$$(\tilde{\mathbf{u}}, \tilde{y}) = (\mathbf{u}, \texttt{randperm}(y))$$

$$(\tilde{\mathbf{u}}, \tilde{y}) \sim p(\mathbf{u})p(y) \Leftrightarrow \tilde{\mathbf{u}} \text{ and } \tilde{y} \text{ independent}$$

Permutation-testing can be used for accurate estimation of the FWER



$\tilde{y}$	$\tilde{u}_1$	$\tilde{u}_2$	$\tilde{u}_3$	$\tilde{u}_4$	$\tilde{u}_5$	$\tilde{u}_6$	$\tilde{u}_7$	$\tilde{u}_8$	$\tilde{u}_9$	$\tilde{u}_{10}$
	1	1	0	0	1	1	0	1	1	1
	1	1	0	1	1	1	0	0	1	0
	1	0	1	0	1	0	0	0	1	1
	1	1	1	0	1	1	1	0	0	1
	0	1	1	0	1	1	0	1	1	0
	1	1	1	0	0	1	0	0	0	1
	1	1	1	0	1	1	0	0	1	1
	1	1	1	0	1	1	1	0	1	1
	1	1	1	0	1	1	0	0	1	1
	1	1	1	0	1	1	0	1	0	0
	1	1	1	0	1	1	0	1	1	1
	1	1	0	1	1	1	0	0	1	1

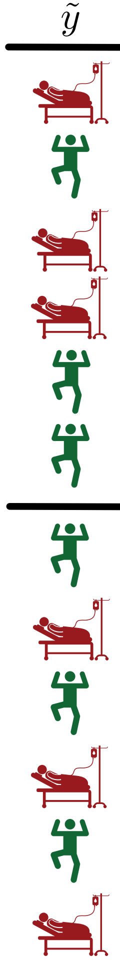
2^p feature interactions

	$z_{\mathcal{S}_1}$	$z_{\mathcal{S}_2}$	$z_{\mathcal{S}_3}$	$\cdots$	$z_{\mathcal{S}_{2^p-1}}$	$z_{\mathcal{S}_{2^p}}$	$\min_{\mathcal{S}} p(z_{\mathcal{S}})$
1	$p^{(1)}(z_{\mathcal{S}_1})$	$p^{(1)}(z_{\mathcal{S}_2})$	$p^{(1)}(z_{\mathcal{S}_3})$	$\cdots$	$p^{(1)}(z_{\mathcal{S}_{2^p-1}})$	$p^{(1)}(z_{\mathcal{S}_{2^p}})$	$p^{(1)}_{min}$

$$(\tilde{\mathbf{u}}, \tilde{y}) = (\mathbf{u}, \texttt{randperm}(y))$$

$$(\tilde{\mathbf{u}}, \tilde{y}) \sim p(\mathbf{u})p(y) \Leftrightarrow \tilde{\mathbf{u}} \text{ and } \tilde{y} \text{ independent}$$

Permutation-testing can be used for accurate estimation of the FWER



$\tilde{u}_1$	$\tilde{u}_2$	$\tilde{u}_3$	$\tilde{u}_4$	$\tilde{u}_5$	$\tilde{u}_6$	$\tilde{u}_7$	$\tilde{u}_8$	$\tilde{u}_9$	$\tilde{u}_{10}$
1	1	0	0	1	1	0	1	1	1
1	1	0	1	1	1	0	0	1	0
1	0	1	0	1	0	0	0	1	1
1	1	1	0	1	1	1	0	0	1
0	1	1	0	1	1	0	1	1	0
1	1	1	0	0	1	0	0	0	1
1	1	1	0	1	1	0	0	1	1
1	1	1	0	1	1	1	0	1	1
1	1	1	0	1	1	0	0	1	1
1	1	1	0	1	1	0	1	0	0
1	1	1	0	1	1	0	1	1	1
1	1	0	1	1	1	0	0	1	1

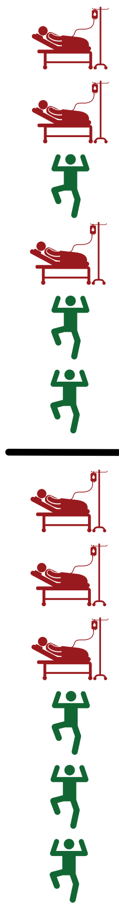
2^p feature interactions

	$z_{\mathcal{S}_1}$	$z_{\mathcal{S}_2}$	$z_{\mathcal{S}_3}$	$\cdots$	$z_{\mathcal{S}_{2^p-1}}$	$z_{\mathcal{S}_{2^p}}$	$\min_{\mathcal{S}} p(z_{\mathcal{S}})$	$\mathrm{FP}(\delta) \neq 0$
1	$p^{(1)}(z_{\mathcal{S}_1})$	$p^{(1)}(z_{\mathcal{S}_2})$	$p^{(1)}(z_{\mathcal{S}_3})$	$\cdots$	$p^{(1)}(z_{\mathcal{S}_{2^p-1}})$	$p^{(1)}(z_{\mathcal{S}_{2^p}})$	$p^{(1)}_{min}$	$\mathbf{1}[p^{(1)}_{min} \leq \delta]$

$$(\tilde{\mathbf{u}}, \tilde{y}) = (\mathbf{u}, \mathtt{randperm}(y))$$

$$(\tilde{\mathbf{u}}, \tilde{y}) \sim p(\mathbf{u})p(y) \Leftrightarrow \tilde{\mathbf{u}} \text{ and } \tilde{y} \text{ independent}$$

Permutation-testing can be used for accurate estimation of the FWER

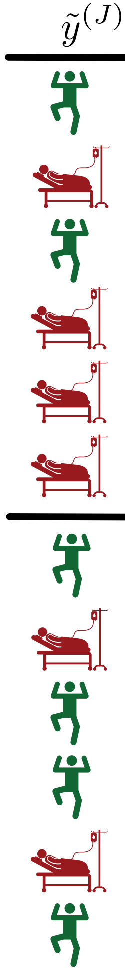


$\tilde{y}^{(2)}$	$\tilde{u}_1$	$\tilde{u}_2$	$\tilde{u}_3$	$\tilde{u}_4$	$\tilde{u}_5$	$\tilde{u}_6$	$\tilde{u}_7$	$\tilde{u}_8$	$\tilde{u}_9$	$\tilde{u}_{10}$
	1	1	0	0	1	1	0	1	1	1
	1	1	0	1	1	1	0	0	1	0
	1	0	1	0	1	0	0	0	1	1
	1	1	1	0	1	1	1	0	0	1
	0	1	1	0	1	1	0	1	1	0
	1	1	1	0	0	1	0	0	0	1
	1	1	1	0	1	1	0	0	1	1
	1	1	1	0	1	1	1	0	1	1
	1	1	1	0	1	1	0	0	1	1
	1	1	1	0	1	1	0	1	0	0
	1	1	1	0	1	1	0	1	1	1
	1	1	0	1	1	1	0	0	1	1

2^p feature interactions

	$z_{\mathcal{S}_1}$	$z_{\mathcal{S}_2}$	$z_{\mathcal{S}_3}$	$\cdots$	$z_{\mathcal{S}_{2^p-1}}$	$z_{\mathcal{S}_{2^p}}$	$\min_{\mathcal{S}} p(z_{\mathcal{S}})$	$\mathrm{FP}(\delta) \neq 0$
1	$p^{(1)}(z_{\mathcal{S}_1})$	$p^{(1)}(z_{\mathcal{S}_2})$	$p^{(1)}(z_{\mathcal{S}_3})$	$\cdots$	$p^{(1)}(z_{\mathcal{S}_{2^p-1}})$	$p^{(1)}(z_{\mathcal{S}_{2^p}})$	$p^{(1)}_{min}$	$\mathbf{1}[p^{(1)}_{min} \leq \delta]$
2	$p^{(2)}(z_{\mathcal{S}_1})$	$p^{(2)}(z_{\mathcal{S}_2})$	$p^{(2)}(z_{\mathcal{S}_3})$	$\cdots$	$p^{(2)}(z_{\mathcal{S}_{2^p-1}})$	$p^{(2)}(z_{\mathcal{S}_{2^p}})$	$p^{(2)}_{min}$	$\mathbf{1}[p^{(2)}_{min} \leq \delta]$

Permutation-testing can be used for accurate estimation of the FWER



$\tilde{u}_1$	$\tilde{u}_2$	$\tilde{u}_3$	$\tilde{u}_4$	$\tilde{u}_5$	$\tilde{u}_6$	$\tilde{u}_7$	$\tilde{u}_8$	$\tilde{u}_9$	$\tilde{u}_{10}$
1	1	0	0	1	1	0	1	1	1
1	1	0	1	1	1	0	0	1	0
1	0	1	0	1	0	0	0	1	1
1	1	1	0	1	1	1	0	0	1
0	1	1	0	1	1	0	1	1	0
1	1	1	0	0	1	0	0	0	1
1	1	1	0	1	1	0	0	1	1
1	1	1	0	1	1	1	0	1	1
1	1	1	0	1	1	0	0	1	1
1	1	1	0	1	1	0	1	0	0
1	1	1	0	1	1	0	1	1	1
1	1	0	1	1	1	0	0	1	1

2^p feature interactions

J random permutations

	$z_{\mathcal{S}_1}$	$z_{\mathcal{S}_2}$	$z_{\mathcal{S}_3}$	$\cdots$	$z_{\mathcal{S}_{2^p-1}}$	$z_{\mathcal{S}_{2^p}}$	$\min_{\mathcal{S}} p(z_{\mathcal{S}})$	$\mathrm{FP}(\delta) \neq 0$
1	$p^{(1)}(z_{\mathcal{S}_1})$	$p^{(1)}(z_{\mathcal{S}_2})$	$p^{(1)}(z_{\mathcal{S}_3})$	$\cdots$	$p^{(1)}(z_{\mathcal{S}_{2^p-1}})$	$p^{(1)}(z_{\mathcal{S}_{2^p}})$	$p^{(1)}_{min}$	$\mathbf{1}[p^{(1)}_{min} \leq \delta]$
2	$p^{(2)}(z_{\mathcal{S}_1})$	$p^{(2)}(z_{\mathcal{S}_2})$	$p^{(2)}(z_{\mathcal{S}_3})$	$\cdots$	$p^{(2)}(z_{\mathcal{S}_{2^p-1}})$	$p^{(2)}(z_{\mathcal{S}_{2^p}})$	$p^{(2)}_{min}$	$\mathbf{1}[p^{(2)}_{min} \leq \delta]$
3	$p^{(3)}(z_{\mathcal{S}_1})$	$p^{(3)}(z_{\mathcal{S}_2})$	$p^{(3)}(z_{\mathcal{S}_3})$	$\cdots$	$p^{(3)}(z_{\mathcal{S}_{2^p-1}})$	$p^{(3)}(z_{\mathcal{S}_{2^p}})$	$p^{(3)}_{min}$	$\mathbf{1}[p^{(3)}_{min} \leq \delta]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
J	$p^{(J)}(z_{\mathcal{S}_1})$	$p^{(J)}(z_{\mathcal{S}_2})$	$p^{(J)}(z_{\mathcal{S}_3})$	$\cdots$	$p^{(J)}(z_{\mathcal{S}_{2^p-1}})$	$p^{(J)}(z_{\mathcal{S}_{2^p}})$	$p^{(J)}_{min}$	$\mathbf{1}[p^{(J)}_{min} \leq \delta]$

$$FWER_{wy}(\delta) = \frac{1}{J}\sum_{i=1}^{J} \mathbf{1}[p^{(i)}_{min} \leq \delta]$$

Permutation-testing can be used for accurate estimation of the FWER

$\tilde{y}^{(J)}$	$\tilde{u}_1$	$\tilde{u}_2$	$\tilde{u}_3$	$\tilde{u}_4$	$\tilde{u}_5$	$\tilde{u}_6$	$\tilde{u}_7$	$\tilde{u}_8$	$\tilde{u}_9$	$\tilde{u}_{10}$
	1	1	0	0	1	1	0	1	1	1
	1	1	0	1	1	1	0	0	1	0
	1	0	1	0	1	0	0	0	1	1
	1	1	1	0	1	1	1	0	0	1
	0	1	1	0	1	1	0	1	1	0
	1	1	1	0	0	1	0	0	0	1
	1	1	1	0	1	1	0	0	1	1
	1	1	1	0	1	1	1	0	1	1
	1	1	1	0	1	1	0	0	1	1
	1	1	1	0	1	1	0	1	0	0
	1	1	1	0	1	1	0	1	1	1
	1	1	0	1	1	1	0	0	1	1

2^p feature interactions

J random permutations

	$z_{\mathcal{S}_1}$	$z_{\mathcal{S}_2}$	$z_{\mathcal{S}_3}$	$\cdots$	$z_{\mathcal{S}_{2^p-1}}$	$z_{\mathcal{S}_{2^p}}$	$\min_{\mathcal{S}} p(z_{\mathcal{S}})$	$\mathrm{FP}(\delta) \neq 0$
1	$p^{(1)}(z_{\mathcal{S}_1})$	$p^{(1)}(z_{\mathcal{S}_2})$	$p^{(1)}(z_{\mathcal{S}_3})$	$\cdots$	$p^{(1)}(z_{\mathcal{S}_{2^p-1}})$	$p^{(1)}(z_{\mathcal{S}_{2^p}})$	$p^{(1)}_{min}$	$\mathbf{1}[p^{(1)}_{min} \leq \delta]$
2	$p^{(2)}(z_{\mathcal{S}_1})$	$p^{(2)}(z_{\mathcal{S}_2})$	$p^{(2)}(z_{\mathcal{S}_3})$	$\cdots$	$p^{(2)}(z_{\mathcal{S}_{2^p-1}})$	$p^{(2)}(z_{\mathcal{S}_{2^p}})$	$p^{(2)}_{min}$	$\mathbf{1}[p^{(2)}_{min} \leq \delta]$
3	$p^{(3)}(z_{\mathcal{S}_1})$	$p^{(3)}(z_{\mathcal{S}_2})$	$p^{(3)}(z_{\mathcal{S}_3})$	$\cdots$	$p^{(3)}(z_{\mathcal{S}_{2^p-1}})$	$p^{(3)}(z_{\mathcal{S}_{2^p}})$	$p^{(3)}_{min}$	$\mathbf{1}[p^{(3)}_{min} \leq \delta]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
J	$p^{(J)}(z_{\mathcal{S}_1})$	$p^{(J)}(z_{\mathcal{S}_2})$	$p^{(J)}(z_{\mathcal{S}_3})$	$\cdots$	$p^{(J)}(z_{\mathcal{S}_{2^p-1}})$	$p^{(J)}(z_{\mathcal{S}_{2^p}})$	$p^{(J)}_{min}$	$\mathbf{1}[p^{(J)}_{min} \leq \delta]$

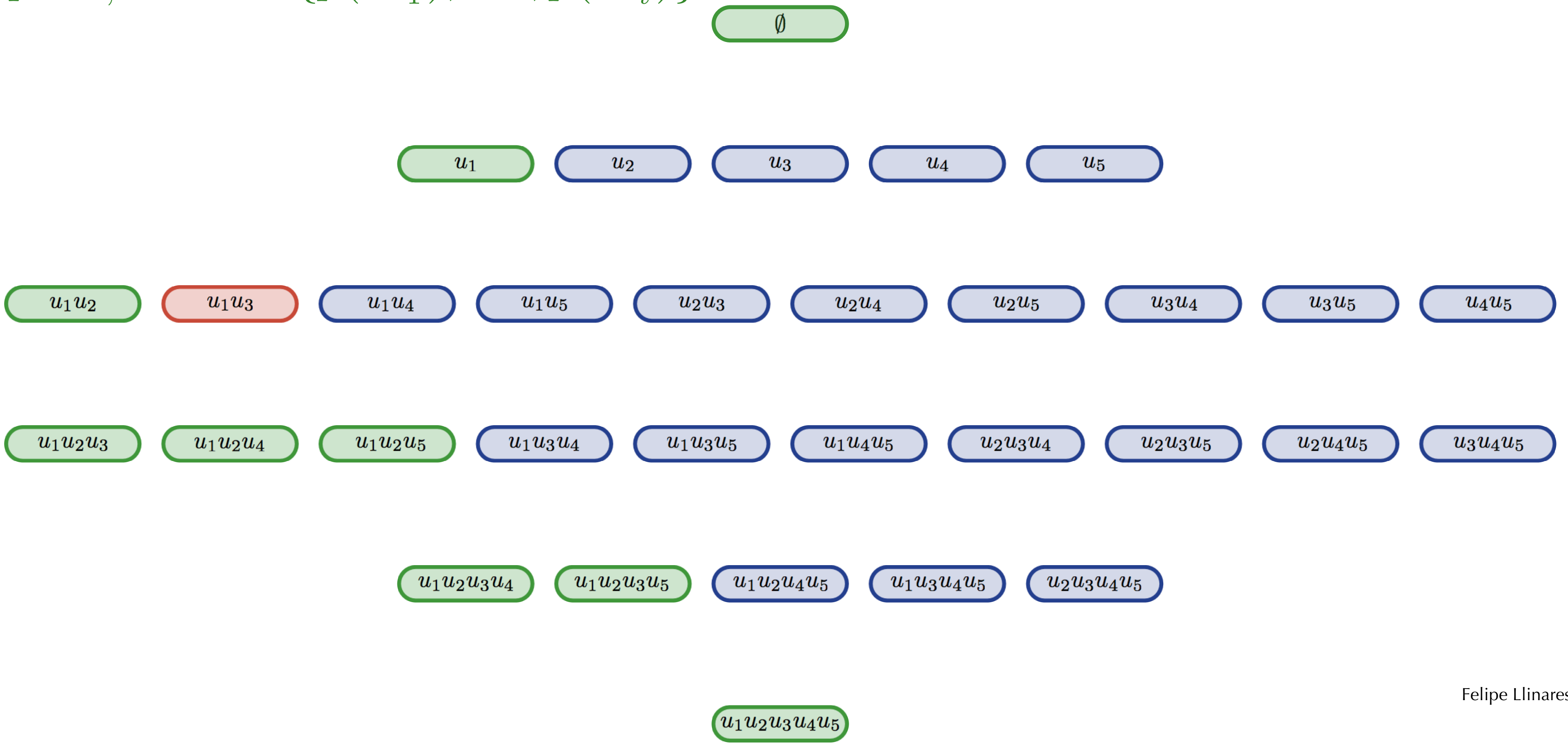
$$FWER_{wy}(\delta) = \frac{1}{J}\sum_{i=1}^{J} \mathbf{1}[p^{(i)}_{min} \leq \delta]$$

Computationally unfeasible for pattern mining!

FastWY (Terada et al., ICBB 2013) uses testability to compute $p_{min}^{(i)}$ efficiently

$z_{\mathcal{S}_1}$	$z_{\mathcal{S}_2}$	$z_{\mathcal{S}_3}$	$\cdots$	$z_{\mathcal{S}_{2^p-1}}$	$z_{\mathcal{S}_{2^p}}$	$\min_{\mathcal{S}} p(z_{\mathcal{S}})$
$p(z_{\mathcal{S}_1})$	$p(z_{\mathcal{S}_2})$	$p(z_{\mathcal{S}_3})$	$\cdots$	$p(z_{\mathcal{S}_{2^p-1}})$	$p(z_{\mathcal{S}_{2^p}})$	p_{min}

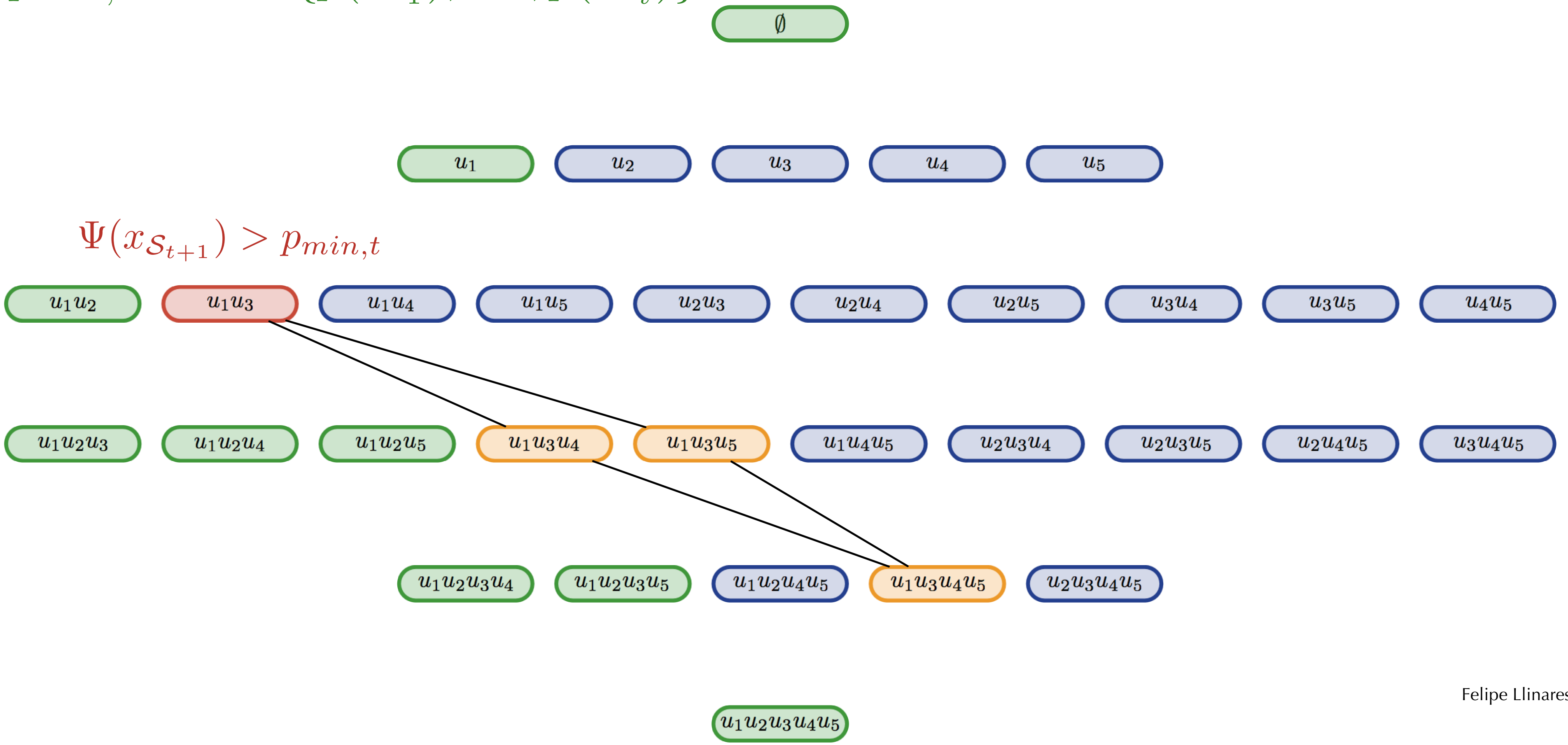
$$p_{min,t} = \min \; \{p(z_{\mathcal{S}_1}), \ldots, p(z_{\mathcal{S}_t})\}$$



FastWY (Terada et al., ICBB 2013) uses testability to compute $p_{min}^{(i)}$ efficiently

$z_{\mathcal{S}_1}$	$z_{\mathcal{S}_2}$	$z_{\mathcal{S}_3}$	$\cdots$	$z_{\mathcal{S}_{2^p-1}}$	$z_{\mathcal{S}_{2^p}}$	$\min_{\mathcal{S}} p(z_{\mathcal{S}})$
$p(z_{\mathcal{S}_1})$	$p(z_{\mathcal{S}_2})$	$p(z_{\mathcal{S}_3})$	$\cdots$	$p(z_{\mathcal{S}_{2^p-1}})$	$p(z_{\mathcal{S}_{2^p}})$	p_{min}

$$p_{min,t} = \min \ \{p(z_{\mathcal{S}_1}), \ldots, p(z_{\mathcal{S}_t})\}$$



Westfall-Young Light (Llinares-López et al., KDD 2015)

Westfall-Young Light (Llinares-López et al., KDD 2015)

- FastWY computes $p_{min}^{(i)}$ for a single permutation, but the entire process must be repeated $J \approx 10{,}000$ times

Westfall-Young Light (Llinares-López et al., KDD 2015)

- FastWY computes $p_{min}^{(i)}$ for a single permutation, but the entire process must be repeated $J \approx 10{,}000$ times

- **<u>Main insight</u>:**
  - Computing $\delta_{wy}^{*} = \max \delta \;\; \text{s.t.} \;\; FWER_{wy}(\delta) \leq \alpha$ is a simpler task than finding $p_{min}^{(i)}$ for all resampled datasets $i = 1, \dots, J$

Westfall-Young Light (Llinares-López et al., KDD 2015)

- FastWY computes $p_{min}^{(i)}$ for a single permutation, but the entire process must be repeated $J \approx 10{,}000$ times

- **Main insight**:
  - Computing $\delta_{wy}^{*} = \max \delta \;\; \text{s.t.} \;\; FWER_{wy}(\delta) \leq \alpha$ is a simpler task than finding $p_{min}^{(i)}$ for all resampled datasets $i = 1, \ldots, J$

WY-Light finds the significance threshold δ_{wy}^{*} directly, processing all J permutations simultaneously and bypassing the need to compute $p_{min}^{(i)}$ for each individual permutation

The Algorithm (Westfall-Young Light)

The Algorithm (Westfall-Young Light)

- **Input:** Feature matrix $U \in \{0,1\}^{n \times p}$, class labels $y \in \{0,1\}^n$, target FWER α, number of permutations J
- **Initialization:**
  1. Compute and store J independent random permutations of the vector of class labels y
  2. Initialize significance threshold δ to 1
  3. Initialize minimum p-value $p_{min}^{(i)}$ for each of the J permutations to 1

The Algorithm (Westfall-Young Light)

- **Input:** Feature matrix $U \in \{0,1\}^{n \times p}$, class labels $y \in \{0,1\}^n$, target FWER α, number of permutations J
- **Initialization:**
  1. Compute and store J independent random permutations of the vector of class labels y
  2. Initialize significance threshold δ to 1
  3. Initialize minimum p-value $p_{min}^{(i)}$ for each of the J permutations to 1
- $\mathbf{DFS}(\emptyset)$

The Algorithm (Westfall-Young Light)

- **Input:** Feature matrix $U \in \{0,1\}^{n \times p}$, class labels $y \in \{0,1\}^n$, target FWER α, number of permutations J
- **Initialization:**
  1. Compute and store J independent random permutations of the vector of class labels y
  2. Initialize significance threshold δ to 1
  3. Initialize minimum p-value $p_{min}^{(i)}$ for each of the J permutations to 1
- $\mathbf{DFS}(\emptyset)$

- $\mathbf{DFS}(\mathcal{S})$:
  1. Update minimum p-value so far for each permutation
  2. Compute lower bound on FWER based on minimum p-values so far
  3. If FWER condition is violated, decrease significance threshold until restored
  4. Continue depth-first search recursively by visiting all testable children of $\mathcal{S}$

The Algorithm (Westfall-Young Light)

- **Input:** Feature matrix $U \in \{0,1\}^{n \times p}$, class labels $y \in \{0,1\}^n$, target FWER α, number of permutations J
- **Initialization:**
  1. Compute and store J independent random permutations of the vector of class labels y
  2. Initialize significance threshold δ to 1
  3. Initialize minimum p-value $p_{min}^{(i)}$ for each of the J permutations to 1
- $\mathbf{DFS}(\emptyset)$
- **Return** $\lfloor \alpha \boldsymbol{J} \rfloor$ **smallest** $p_{min}^{(i)}$

  - $\mathbf{DFS}(\mathcal{S})$**:**
    1. Update minimum p-value so far for each permutation
    2. Compute lower bound on FWER based on minimum p-values so far
    3. If FWER condition is violated, decrease significance threshold until restored
    4. Continue depth-first search recursively by visiting all testable children of $\mathcal{S}$

Key features of Westfall-Young Light

i. Guaranteed to return the same solution δ^*_{wy} as FastWY

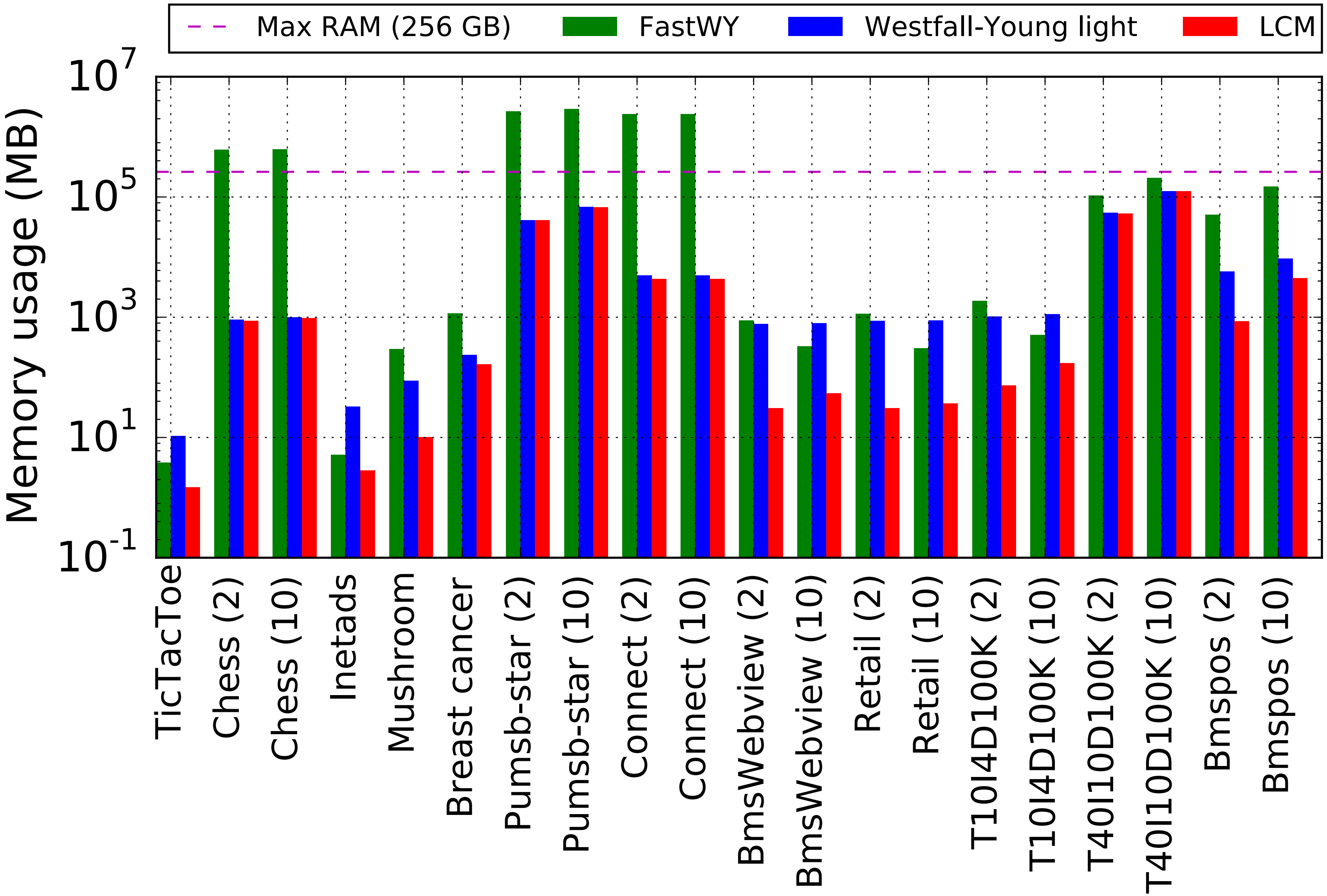
ii. It only needs to enumerate patterns once, instead of one time per permutation

iii. It does not require additional memory usage to compensate for the need to repeat pattern enumeration

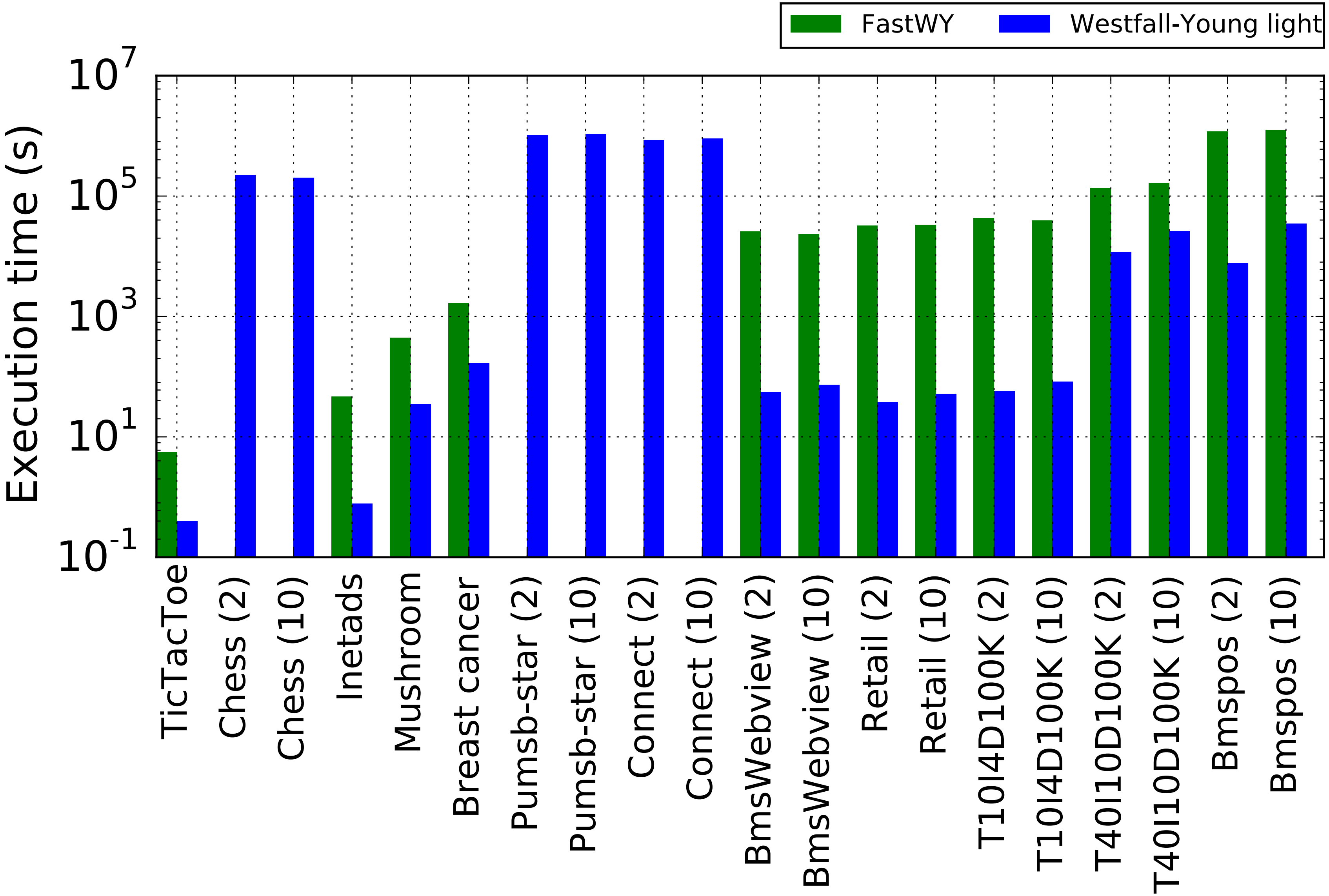
iv. Only needs to compute exactly the smallest $[\alpha J]$ $p^{(i)}_{min}$, greatly reducing the number of patterns that need to be enumerated to find the solution

v. The computation of p-values is shared across all J permutations

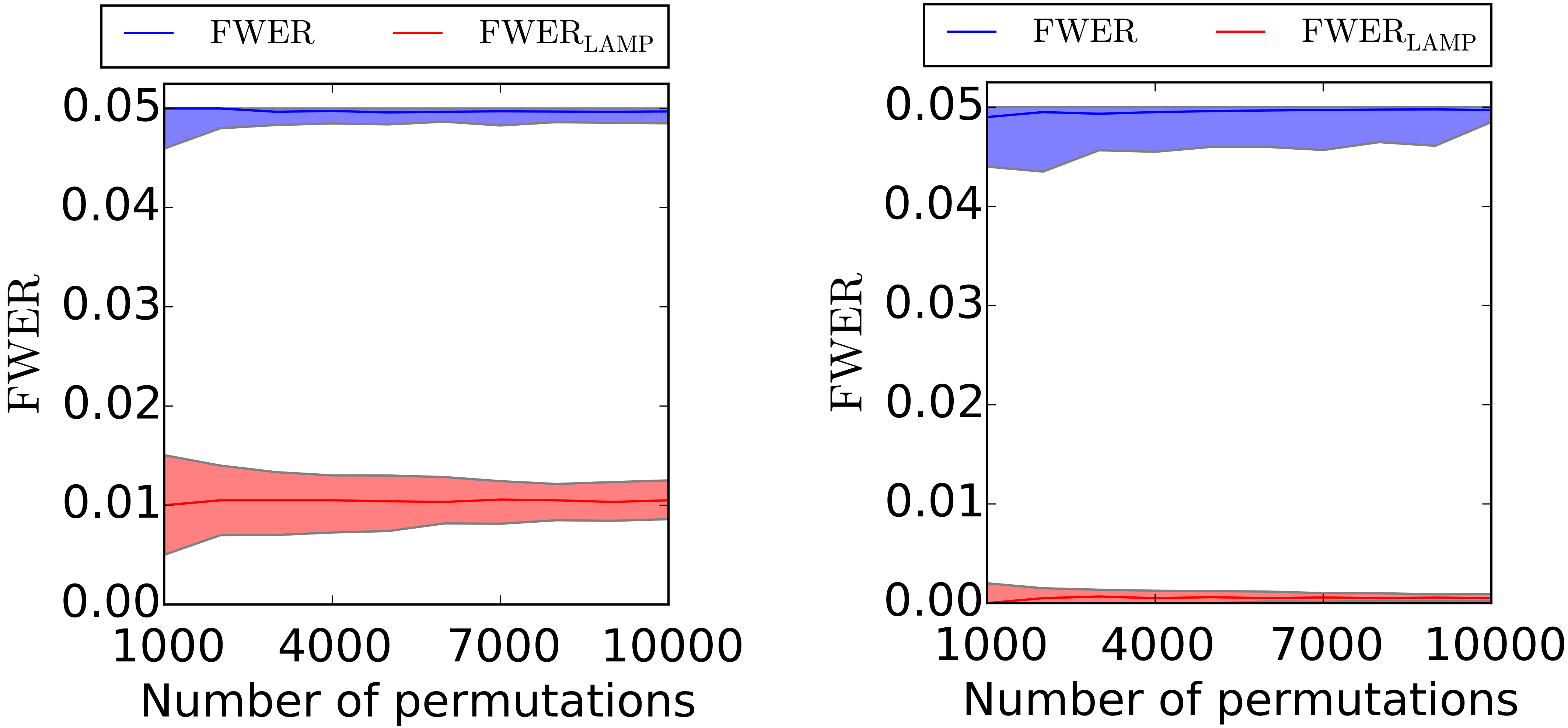
WY-Light improves over the state-of-the-art in terms of memory



WY-Light improves over the state-of-the-art in terms of runtime



WY-Light accurately estimates the FWER



Correcting for an observed categorical covariate

Ignoring covariate factors might lead to many spurious discoveries

p features

n samples

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	$z_{\mathcal{S}_1}$	$z_{\mathcal{S}_2}$
	1	1	0	0	1	1	0	1	1	1	0	1
	1	1	0	1	1	1	0	0	1	0	0	0
	1	0	1	0	1	0	0	0	1	1	0	0
	1	1	1	0	1	1	1	0	0	1	1	0
	0	1	1	0	1	1	0	1	1	0	0	0
	1	1	1	0	0	1	0	0	0	1	0	0
	1	1	1	0	1	1	0	0	1	1	1	1
	1	1	1	0	1	1	1	0	1	1	1	1
	1	1	1	0	1	1	0	0	1	1	1	1
	1	1	1	0	1	1	0	1	0	0	1	0
	1	1	1	0	1	1	0	1	1	1	1	1
	1	1	0	1	1	1	0	0	1	1	0	1

Ignoring covariate factors might lead to many spurious discoveries

p features

n samples

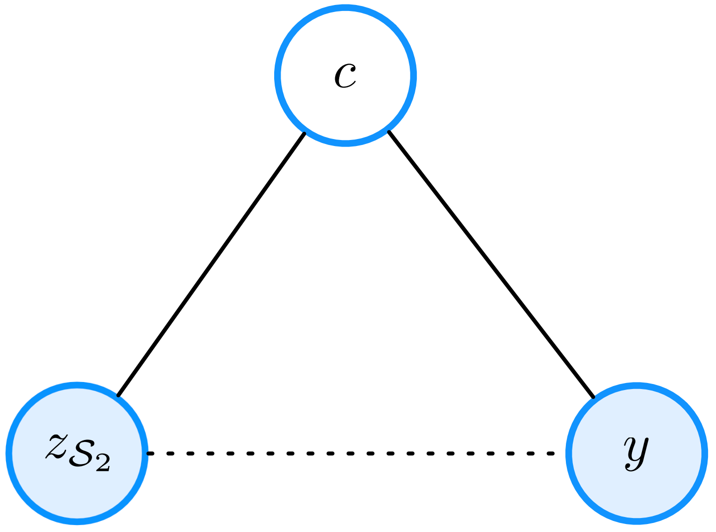
y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	c	$z_{\mathcal{S}_1}$	$z_{\mathcal{S}_2}$
	1	1	0	0	1	1	0	1	1	1		0	1
	1	1	0	1	1	1	0	0	1	0		0	0
	1	0	1	0	1	0	0	0	1	1		0	0
	1	1	1	0	1	1	1	0	0	1		1	0
	0	1	1	0	1	1	0	1	1	0		0	0
	1	1	1	0	0	1	0	0	0	1		0	0
	1	1	1	0	1	1	0	0	1	1		1	1
	1	1	1	0	1	1	1	0	1	1		1	1
	1	1	1	0	1	1	0	0	1	1		1	1
	1	1	1	0	1	1	0	1	0	0		1	0
	1	1	1	0	1	1	0	1	1	1		1	1
	1	1	0	1	1	1	0	0	1	1		0	1

Ignoring covariate factors might lead to many spurious discoveries

p features

n samples

u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	$z_{\mathcal{S}_1}$	$z_{\mathcal{S}_2}$
1	1	0	0	1	1	0	1	1	1	0	1
1	1	0	1	1	1	0	0	1	0	0	0
1	0	1	0	1	0	0	0	1	1	0	0
1	1	1	0	1	1	1	0	0	1	1	0
0	1	1	0	1	1	0	1	1	0	0	0
1	1	1	0	0	1	0	0	0	1	0	0
1	1	1	0	1	1	0	0	1	1	1	1
1	1	1	0	1	1	1	0	1	1	1	1
1	1	1	0	1	1	0	0	1	1	1	1
1	1	1	0	1	1	0	1	0	0	1	0
1	1	1	0	1	1	0	1	1	1	1	1
1	1	0	1	1	1	0	0	1	1	0	1



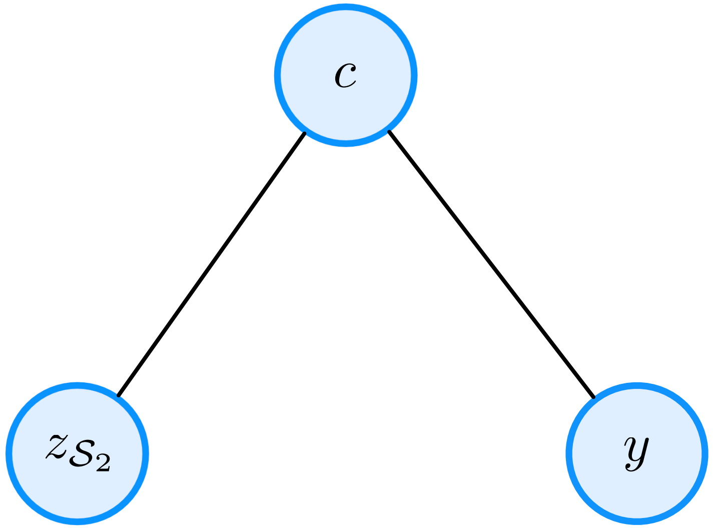
- $z_{\mathcal{S}_2}$ marginally associated to y

Ignoring covariate factors might lead to many spurious discoveries

p features

n samples

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	c	$z_{\mathcal{S}_1}$	$z_{\mathcal{S}_2}$
	1	1	0	0	1	1	0	1	1	1		0	1
	1	1	0	1	1	1	0	0	1	0		0	0
	1	0	1	0	1	0	0	0	1	1		0	0
	1	1	1	0	1	1	1	0	0	1		1	0
	0	1	1	0	1	1	0	1	1	0		0	0
	1	1	1	0	0	1	0	0	0	1		0	0
	1	1	1	0	1	1	0	0	1	1		1	1
	1	1	1	0	1	1	1	0	1	1		1	1
	1	1	1	0	1	1	0	0	1	1		1	1
	1	1	1	0	1	1	0	1	0	0		1	0
	1	1	1	0	1	1	0	1	1	1		1	1
	1	1	0	1	1	1	0	0	1	1		0	1



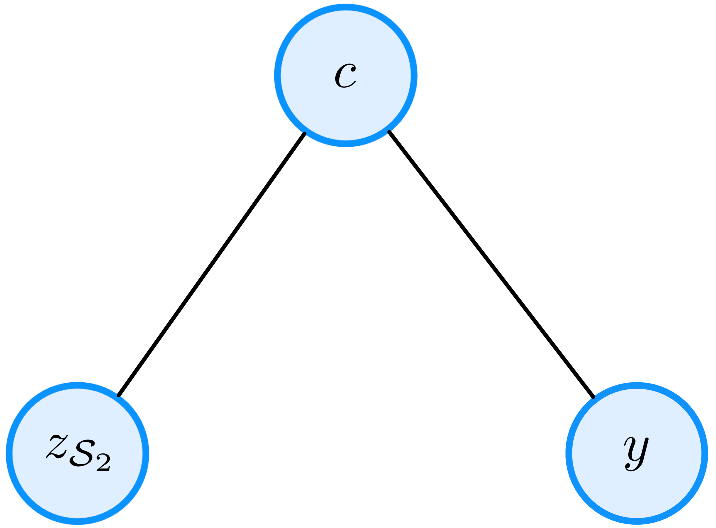
- $z_{\mathcal{S}_2}$ marginally associated to y
- $z_{\mathcal{S}_2}$ independent of y given c

Ignoring covariate factors might lead to many spurious discoveries

p features

n samples

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	c	$z_{\mathcal{S}_1}$	$z_{\mathcal{S}_2}$
	1	1	0	0	1	1	0	1	1	1		0	1
	1	1	0	1	1	1	0	0	1	0		0	0
	1	0	1	0	1	0	0	0	1	1		0	0
	1	1	1	0	1	1	1	0	0	1		1	0
	0	1	1	0	1	1	0	1	1	0		0	0
	1	1	1	0	0	1	0	0	0	1		0	0
	1	1	1	0	1	1	0	0	1	1		1	1
	1	1	1	0	1	1	1	0	1	1		1	1
	1	1	1	0	1	1	0	0	1	1		1	1
	1	1	1	0	1	1	0	1	0	0		1	0
	1	1	1	0	1	1	0	1	1	1		1	1
	1	1	0	1	1	1	0	0	1	1		0	1



- $z_{\mathcal{S}_2}$ marginally associated to y
- $z_{\mathcal{S}_2}$ independent of y given c
- We treat $z_{\mathcal{S}_2}$ as a false positive!

Finding significant combinations of features in the presence of categorical covariates

Laetitia Papaxanthos*, Felipe Llinares-López*, Dean Bodenham, Karsten Borgwardt
Machine Learning and Computational Biology Lab
D-BSSE, ETH Zurich

*Equally contributing authors.

Accepted at NIPS 2016

Goal: Propose a significant pattern mining approach that allows correcting for a categorical covariate

Conditional association testing with the Cochran-Mantel-Haenszel (CMH) test

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	c	$z_{\mathcal{S}_1}$	$z_{\mathcal{S}_2}$
	1	1	0	0	1	1	0	1	1	1		0	1
	1	1	0	1	1	1	0	0	1	0		0	0
	1	0	1	0	1	0	0	0	1	1		0	0
	1	1	1	0	1	1	1	0	0	1		1	0
	0	1	1	0	1	1	0	1	1	0		0	0
	1	1	1	0	0	1	0	0	0	1		0	0
	1	1	1	0	1	1	0	0	1	1		1	1
	1	1	1	0	1	1	1	0	1	1		1	1
	1	1	1	0	1	1	0	0	1	1		1	1
	1	1	1	0	1	1	0	1	0	0		1	0
	1	1	1	0	1	1	0	1	1	1		1	1
	1	1	0	1	1	1	0	0	1	1		0	1

	$z_{\mathcal{S}_1} = 1$	$z_{\mathcal{S}_1} = 0$	
$y = 1$	5	1	6
$y = 0$	1	5	6
	6	6	12

Conditional association testing with the Cochran-Mantel-Haenszel (CMH) test

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	c	$z_{\mathcal{S}_1}$	$z_{\mathcal{S}_2}$
	1	1	0	0	1	1	0	1	1	1		0	1
	1	1	0	1	1	1	0	0	1	0		0	0
	1	0	1	0	1	0	0	0	1	1		0	0
	1	1	1	0	1	1	1	0	0	1		1	0
	0	1	1	0	1	1	0	1	1	0		0	0
	1	1	1	0	0	1	0	0	0	1		0	0
	1	1	1	0	1	1	0	0	1	1		1	1
	1	1	1	0	1	1	1	0	1	1		1	1
	1	1	1	0	1	1	0	0	1	1		1	1
	1	1	1	0	1	1	0	1	0	0		1	0
	1	1	1	0	1	1	0	1	1	1		1	1
	1	1	0	1	1	1	0	0	1	1		0	1

	$z_{\mathcal{S}_1}=1$	$z_{\mathcal{S}_1}=0$	
$y=1$	5	1	6
$y=0$	1	5	6
	6	6	12

	$z_{\mathcal{S}_1}=1$	$z_{\mathcal{S}_1}=0$	
$y=1$	4	1	5
$y=0$	0	1	1
	4	2	6

	$z_{\mathcal{S}_1}=1$	$z_{\mathcal{S}_1}=0$	
$y=1$	1	0	1
$y=0$	1	4	5
	2	4	6

Conditional association testing with the Cochran-Mantel-Haenszel (CMH) test

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	c	$z_{\mathcal{S}_1}$	$z_{\mathcal{S}_2}$
	1	1	0	0	1	1	0	1	1	1		0	1
	1	1	0	1	1	1	0	0	1	0		0	0
	1	0	1	0	1	0	0	0	1	1		0	0
	1	1	1	0	1	1	1	0	0	1		1	0
	0	1	1	0	1	1	0	1	1	0		0	0
	1	1	1	0	0	1	0	0	0	1		0	0
	1	1	1	0	1	1	0	0	1	1		1	1
	1	1	1	0	1	1	1	0	1	1		1	1
	1	1	1	0	1	1	0	0	1	1		1	1
	1	1	1	0	1	1	0	1	0	0		1	0
	1	1	1	0	1	1	0	1	1	1		1	1
	1	1	0	1	1	1	0	0	1	1		0	1

	$z_{\mathcal{S}_1}=1$	$z_{\mathcal{S}_1}=0$	
$y=1$	5	1	6
$y=0$	1	5	6
	6	6	12

	$z_{\mathcal{S}_1}=1$	$z_{\mathcal{S}_1}=0$	
$y=1$	4	1	5
$y=0$	0	1	1
	4	2	6

	$z_{\mathcal{S}_1}=1$	$z_{\mathcal{S}_1}=0$	
$y=1$	1	0	1
$y=0$	1	4	5
	2	4	6

$$p_{cmh}(z_{\mathcal{S}_1}) = 0.029$$

Conditional association testing with the Cochran-Mantel-Haenszel (CMH) test

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	c	$z_{\mathcal{S}_1}$	$z_{\mathcal{S}_2}$
	1	1	0	0	1	1	0	1	1	1		0	1
	1	1	0	1	1	1	0	0	1	0		0	0
	1	0	1	0	1	0	0	0	1	1		0	0
	1	1	1	0	1	1	1	0	0	1		1	0
	0	1	1	0	1	1	0	1	1	0		0	0
	1	1	1	0	0	1	0	0	0	1		0	0
	1	1	1	0	1	1	0	0	1	1		1	1
	1	1	1	0	1	1	1	0	1	1		1	1
	1	1	1	0	1	1	0	0	1	1		1	1
	1	1	1	0	1	1	0	1	0	0		1	0
	1	1	1	0	1	1	0	1	1	1		1	1
	1	1	0	1	1	1	0	0	1	1		0	1

	$z_{\mathcal{S}_2}=1$	$z_{\mathcal{S}_2}=0$	
$y=1$	5	1	6
$y=0$	1	5	6
	6	6	12

Conditional association testing with the Cochran-Mantel-Haenszel (CMH) test

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	c	$z_{\mathcal{S}_1}$	$z_{\mathcal{S}_2}$
	1	1	0	0	1	1	0	1	1	1		0	1
	1	1	0	1	1	1	0	0	1	0		0	0
	1	0	1	0	1	0	0	0	1	1		0	0
	1	1	1	0	1	1	1	0	0	1		1	0
	0	1	1	0	1	1	0	1	1	0		0	0
	1	1	1	0	0	1	0	0	0	1		0	0
	1	1	1	0	1	1	0	0	1	1		1	1
	1	1	1	0	1	1	1	0	1	1		1	1
	1	1	1	0	1	1	0	0	1	1		1	1
	1	1	1	0	1	1	0	1	0	0		1	0
	1	1	1	0	1	1	0	1	1	1		1	1
	1	1	0	1	1	1	0	0	1	1		0	1

	$z_{\mathcal{S}_2}=1$	$z_{\mathcal{S}_2}=0$	
$y=1$	5	1	6
$y=0$	1	5	6
	6	6	12

	$z_{\mathcal{S}_2}=1$	$z_{\mathcal{S}_2}=0$	
$y=1$	5	0	5
$y=0$	1	0	1
	6	0	6

	$z_{\mathcal{S}_2}=1$	$z_{\mathcal{S}_2}=0$	
$y=1$	0	1	1
$y=0$	0	5	5
	0	6	6

Conditional association testing with the Cochran-Mantel-Haenszel (CMH) test

y	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	c	$z_{\mathcal{S}_1}$	$z_{\mathcal{S}_2}$
	1	1	0	0	1	1	0	1	1	1		0	1
	1	1	0	1	1	1	0	0	1	0		0	0
	1	0	1	0	1	0	0	0	1	1		0	0
	1	1	1	0	1	1	1	0	0	1		1	0
	0	1	1	0	1	1	0	1	1	0		0	0
	1	1	1	0	0	1	0	0	0	1		0	0
	1	1	1	0	1	1	0	0	1	1		1	1
	1	1	1	0	1	1	1	0	1	1		1	1
	1	1	1	0	1	1	0	0	1	1		1	1
	1	1	1	0	1	1	0	1	0	0		1	0
	1	1	1	0	1	1	0	1	1	1		1	1
	1	1	0	1	1	1	0	0	1	1		0	1

	$z_{\mathcal{S}_2} = 1$	$z_{\mathcal{S}_2} = 0$	
$y = 1$	5	1	6
$y = 0$	1	5	6
	6	6	12

	$z_{\mathcal{S}_2} = 1$	$z_{\mathcal{S}_2} = 0$	
$y = 1$	5	0	5
$y = 0$	1	0	1
	6	0	6

	$z_{\mathcal{S}_2} = 1$	$z_{\mathcal{S}_2} = 0$	
$y = 1$	0	1	1
$y = 0$	0	5	5
	0	6	6

$$p_{cmh}(z_{\mathcal{S}_2}) = 1.000$$

Significant pattern mining using the Cochran-Mantel-Haenszel test

Significant pattern mining using the Cochran-Mantel-Haenszel test

1. **Does the CMH test have a minimum attainable p-value?**

Significant pattern mining using the Cochran-Mantel-Haenszel test

1. Does the CMH test have a minimum attainable p-value?

Proposition 1: The CMH test has a minimum attainable p-value $\Psi_{cmh}(\mathcal{S})$, which can be computed in $O(k)$ time as a function of the margins of the k contingency tables.

Significant pattern mining using the Cochran-Mantel-Haenszel test

1. Does the CMH test have a minimum attainable p-value?

Proposition 1: The CMH test has a minimum attainable p-value $\Psi_{cmh}(\mathcal{S})$, which can be computed in $O(k)$ time as a function of the margins of the k contingency tables.

- Closed-form expression computable in $O(k)$ time
- Multivariate function: $\Psi_{cmh}(\mathcal{S}) = \Psi_{cmh}\left(x_{\mathcal{S},1}, x_{\mathcal{S},2}, \dots, x_{\mathcal{S},k}\right) = \Psi_{cmh}(\boldsymbol{x}_{\mathcal{S}})$

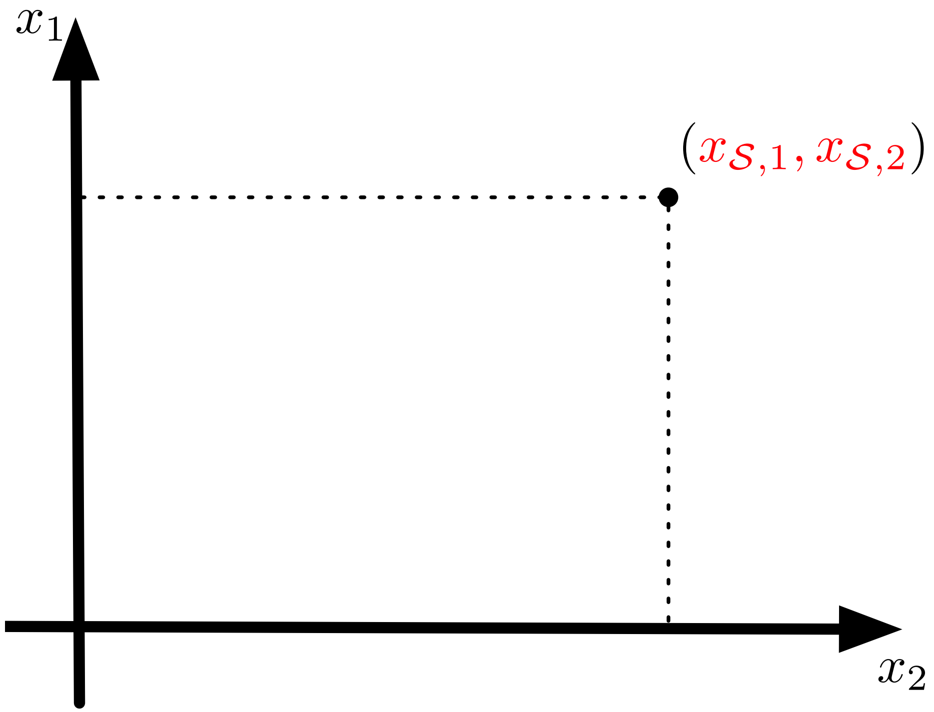
Significant pattern mining using the Cochran-Mantel-Haenszel test

Significant pattern mining using the Cochran-Mantel-Haenszel test

2. **Is the resulting minimum attainable p-value $\Psi_{cmh}(x_{\mathcal{S},1}, x_{\mathcal{S},2}, \dots, x_{\mathcal{S},k})$ function monotonically decreasing?**

Significant pattern mining using the Cochran-Mantel-Haenszel test

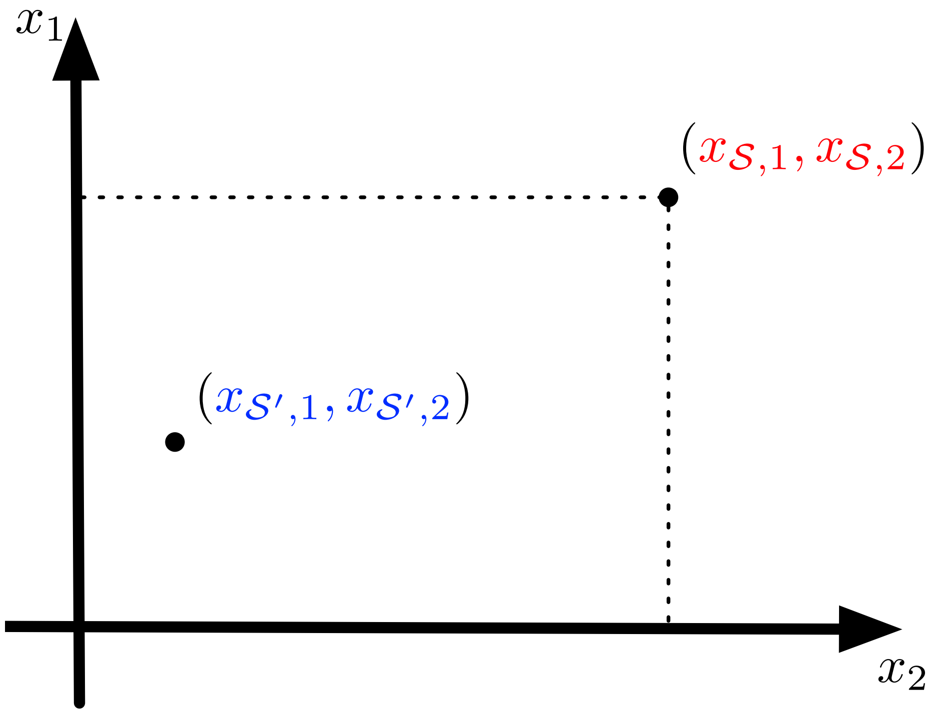
2. Is the resulting minimum attainable p-value $\Psi_{cmh}(x_{\mathcal{S},1}, x_{\mathcal{S},2}, \ldots, x_{\mathcal{S},k})$ function monotonically decreasing?



$\Psi_{cmh}(x_{\mathcal{S},1}, x_{\mathcal{S},2})$

Significant pattern mining using the Cochran-Mantel-Haenszel test

2. Is the resulting minimum attainable p-value $\Psi_{cmh}(x_{\mathcal{S},1}, x_{\mathcal{S},2}, \ldots, x_{\mathcal{S},k})$ function monotonically decreasing?



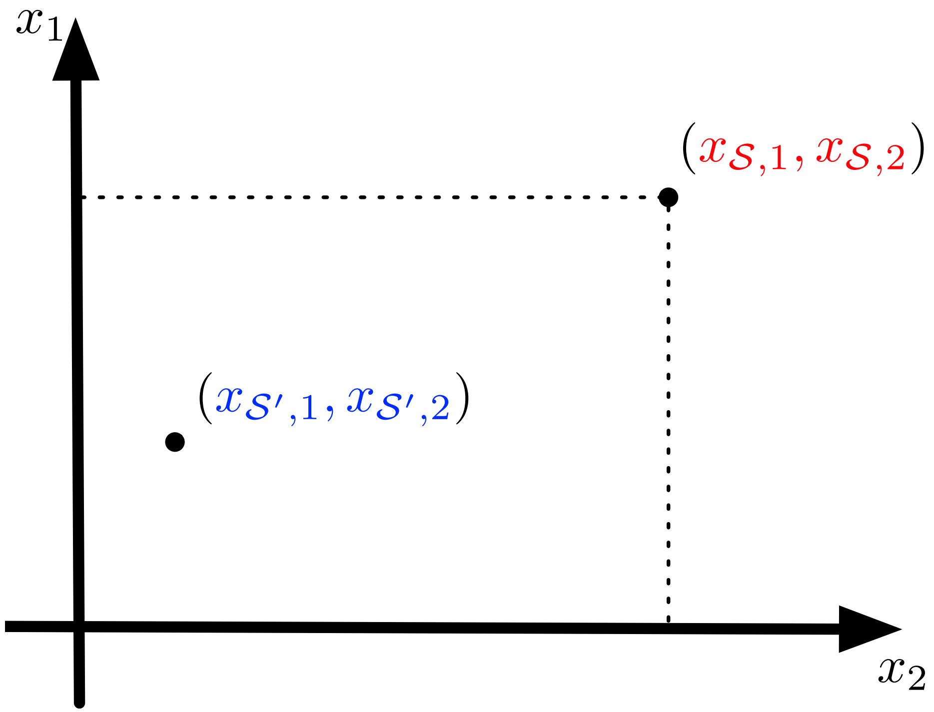
$$x_{\mathcal{S}',j} \leq x_{\mathcal{S},j}, \quad \forall j = 1, \ldots, k$$

$$\Downarrow \; ?$$

$$\Psi_{cmh}(x_{\mathcal{S},1}, x_{\mathcal{S},2}) \leq \Psi_{cmh}(x_{\mathcal{S}',1}, x_{\mathcal{S}',2})$$

Significant pattern mining using the Cochran-Mantel-Haenszel test

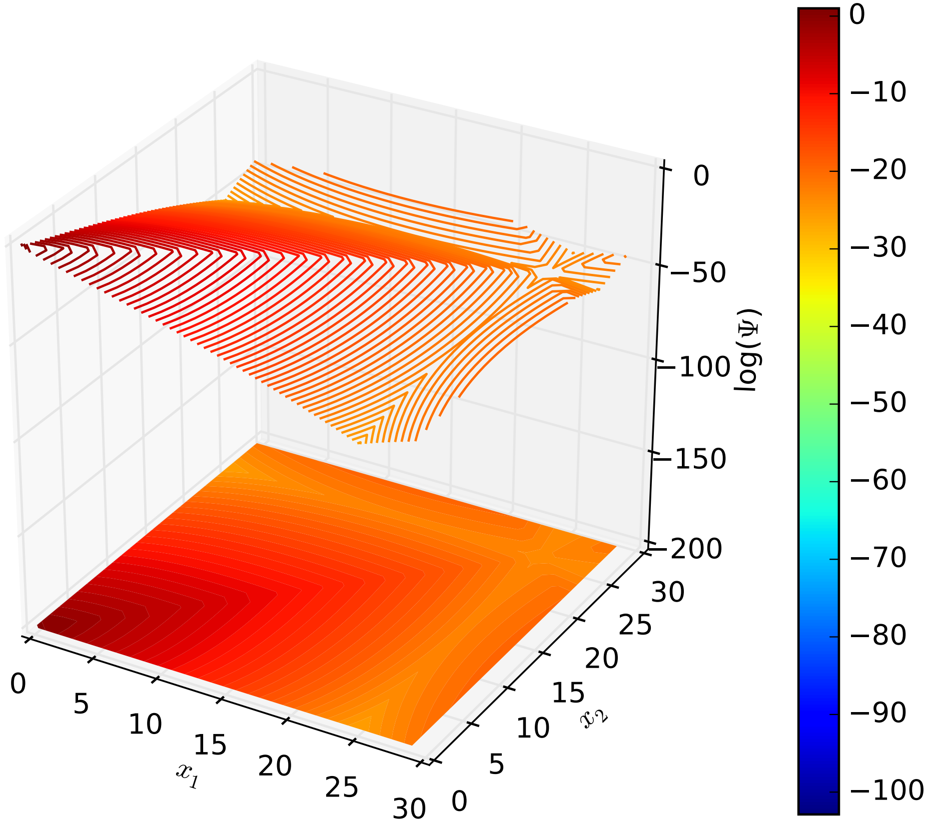
2. Is the resulting minimum attainable p-value $\Psi_{cmh}(x_{\mathcal{S},1}, x_{\mathcal{S},2}, \ldots, x_{\mathcal{S},k})$ function monotonically decreasing?



$$x_{\mathcal{S}',j} \leq x_{\mathcal{S},j}, \quad \forall j = 1, \ldots, k$$

$$\Downarrow \; ?$$

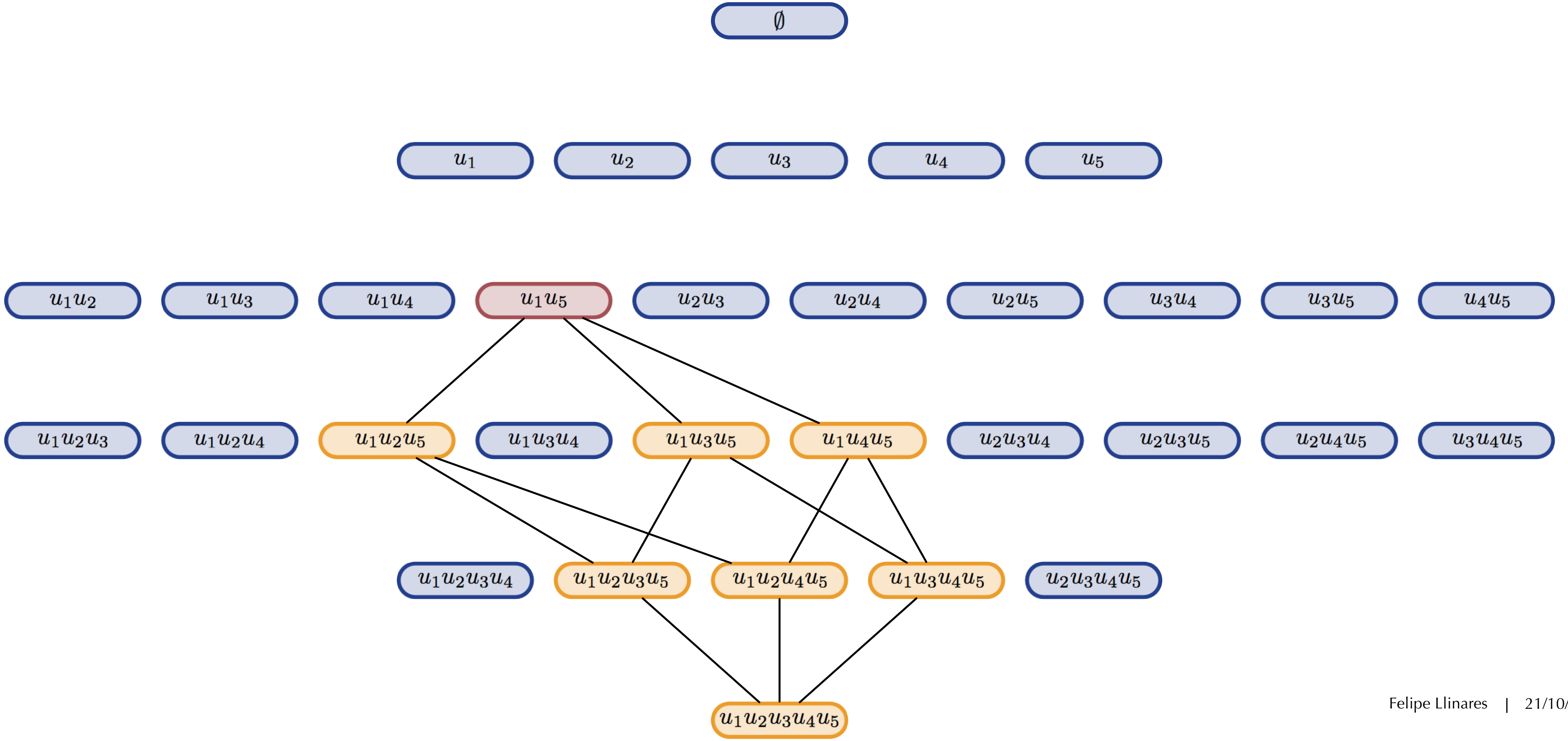
$$\Psi_{cmh}(x_{\mathcal{S},1}, x_{\mathcal{S},2}) \leq \Psi_{cmh}(x_{\mathcal{S}',1}, x_{\mathcal{S}',2})$$



NO!

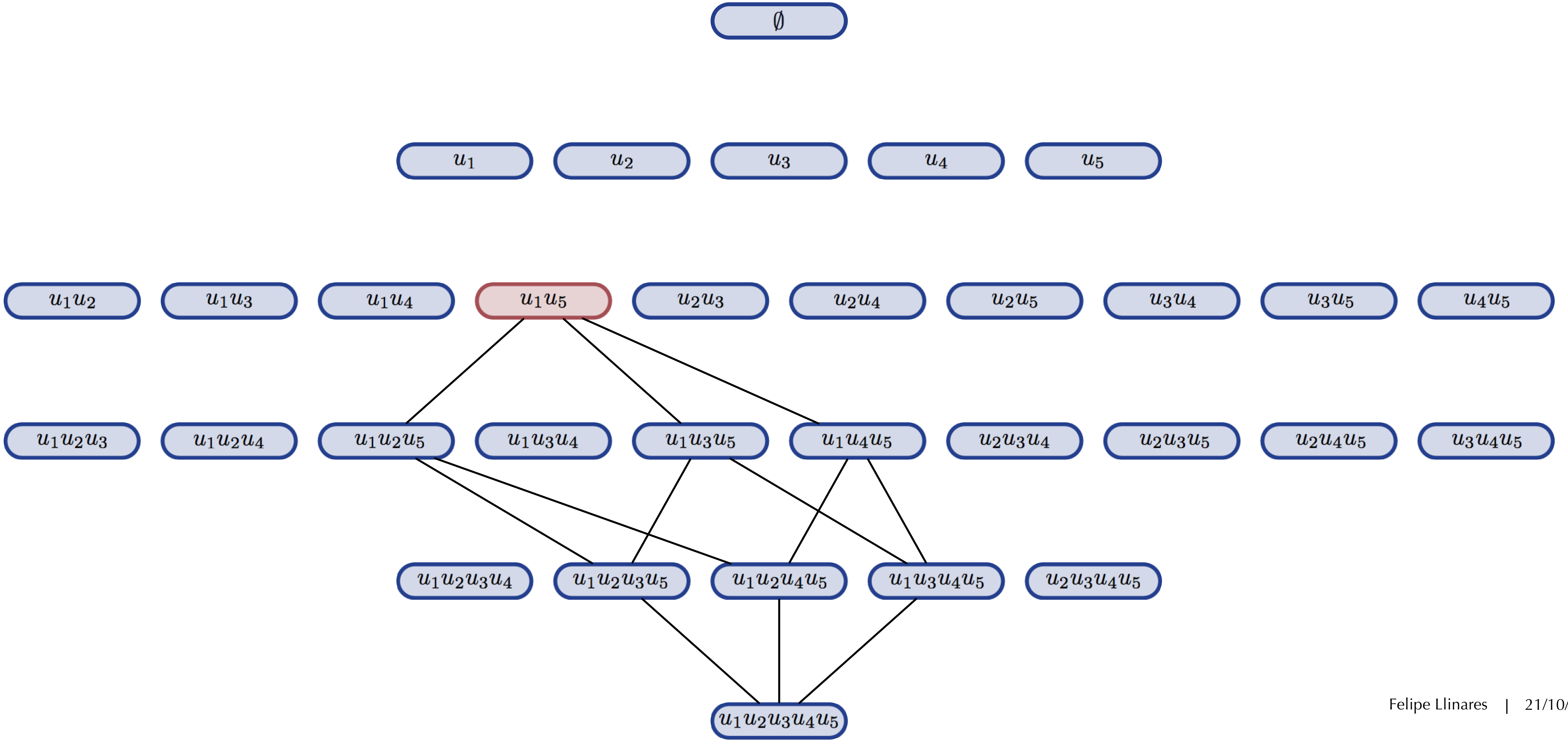
The lack of monotonicity of $\Psi_{cmh}(x_S)$ makes the CMH test incompatible with LAMP

$$\left.\begin{array}{l} 1)\ x_{\mathcal{S}} \geq x_{\mathcal{S}'} \Rightarrow \Psi(x_{\mathcal{S}}) \leq \Psi(x_{\mathcal{S}'}) \\ 2)\ \mathcal{S} \subseteq \mathcal{S}' \Rightarrow x_{\mathcal{S}} \geq x_{\mathcal{S}'} \end{array}\right\} \text{If } \mathcal{S} \subseteq \mathcal{S}', \Psi(x_{\mathcal{S}}) > \delta \Rightarrow \Psi(x_{\mathcal{S}'}) > \delta$$



The lack of monotonicity of $\Psi_{cmh}(x_{\mathcal{S}})$ makes the CMH test incompatible with LAMP

$$\left.\begin{array}{l} 1)\ x_{\mathcal{S}} \geq x_{\mathcal{S}'} \not\Rightarrow \Psi(x_{\mathcal{S}}) \leq \Psi(x_{\mathcal{S}'}) \\ 2)\ \mathcal{S} \subseteq \mathcal{S}' \Rightarrow x_{\mathcal{S}} \geq x_{\mathcal{S}'} \end{array}\right\} \text{If } \mathcal{S} \subseteq \mathcal{S}', \Psi(x_{\mathcal{S}}) > \delta \not\Rightarrow \Psi(x_{\mathcal{S}'}) > \delta$$

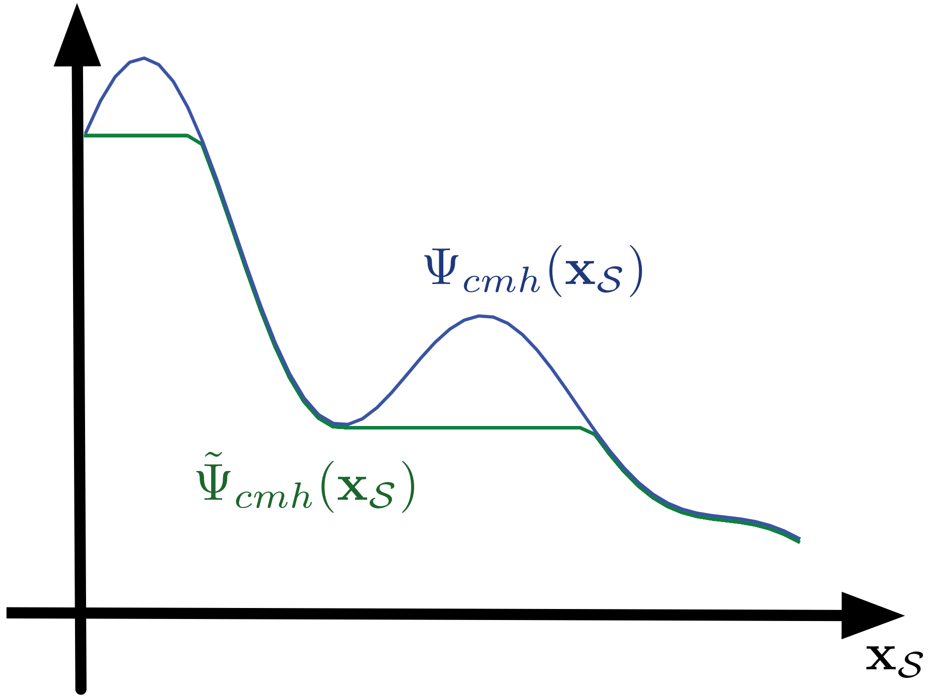


Pruning the search space with a monotonic surrogate of $\Psi_{cmh}(x_s)$

Pruning the search space with a monotonic surrogate of $\Psi_{cmh}(x_S)$

- **Definition:** The *lower envelope* of $\Psi_{cmh}(x_S)$ is defined as:

$$\tilde{\Psi}_{cmh}(\mathbf{x}_{\mathcal{S}}) = \min_{\mathbf{x}_{\mathcal{S}'} \leq \mathbf{x}_{\mathcal{S}}} \Psi_{cmh}(\mathbf{x}_{\mathcal{S}'})$$

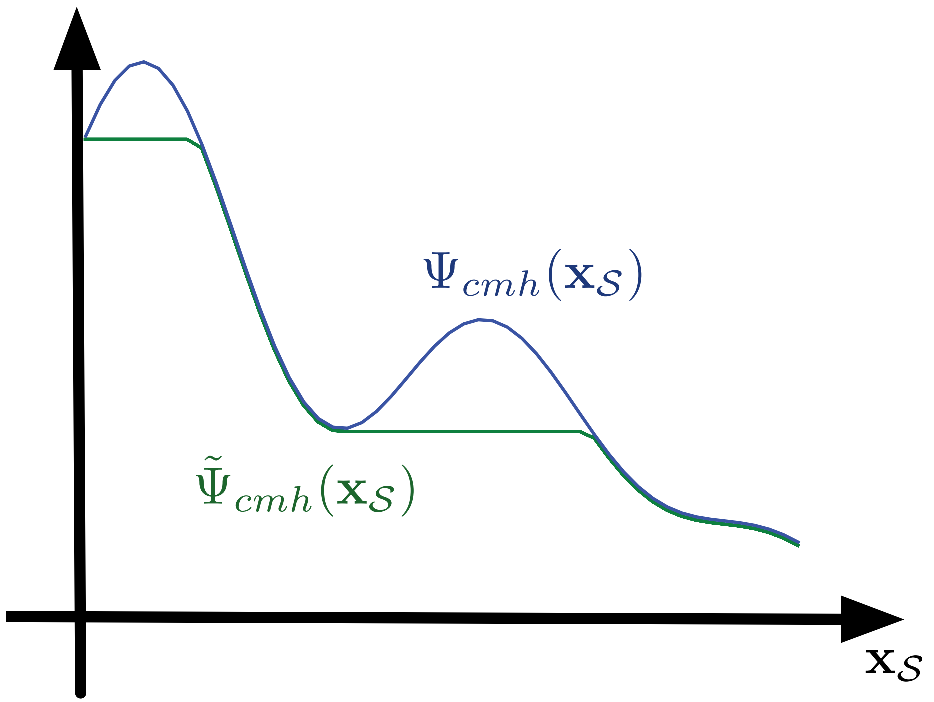


Pruning the search space with a monotonic surrogate of $\Psi_{cmh}(x_S)$

- **Definition:** The *lower envelope* of $\Psi_{cmh}(x_S)$ is defined as:

$$\tilde{\Psi}_{cmh}(\mathbf{x}_{\mathcal{S}}) = \min_{\mathbf{x}_{\mathcal{S}'} \leq \mathbf{x}_{\mathcal{S}}} \Psi_{cmh}(\mathbf{x}_{\mathcal{S}'})$$

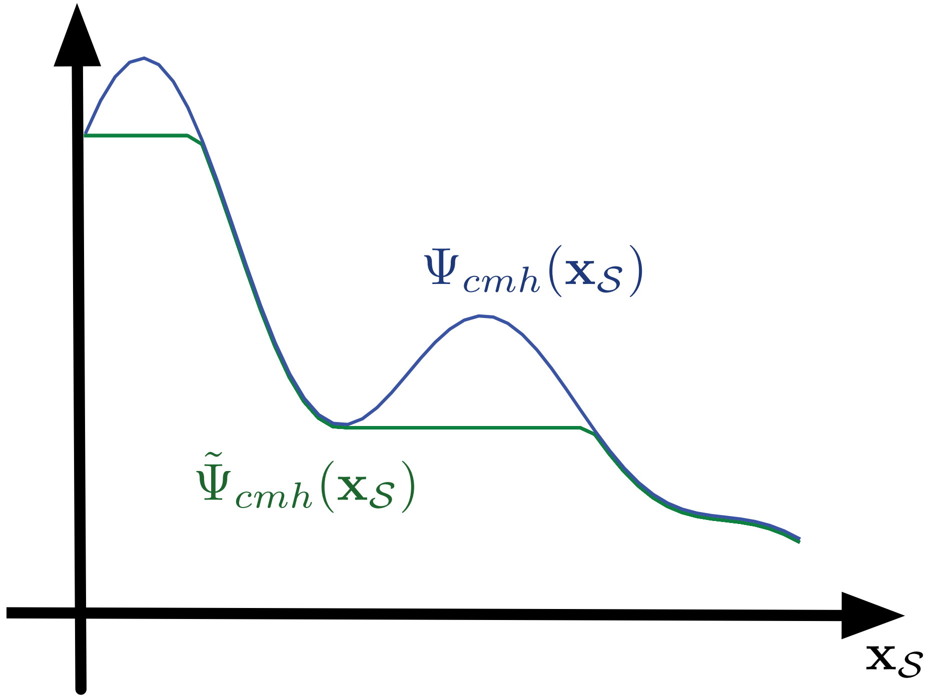
- By construction:
  1. $\tilde{\Psi}_{cmh}(x_S)$ is monotonically decreasing
  2. $\tilde{\Psi}_{cmh}(x_S)$ is a lower bound of $\Psi_{cmh}(x_S)$



Pruning the search space with a monotonic surrogate of $\Psi_{cmh}(x_S)$

- **Definition:** The *lower envelope* of $\Psi_{cmh}(x_S)$ is defined as:

$$\tilde{\Psi}_{cmh}(\mathbf{x}_{\mathcal{S}}) = \min_{\mathbf{x}_{\mathcal{S}'} \leq \mathbf{x}_{\mathcal{S}}} \Psi_{cmh}(\mathbf{x}_{\mathcal{S}'})$$

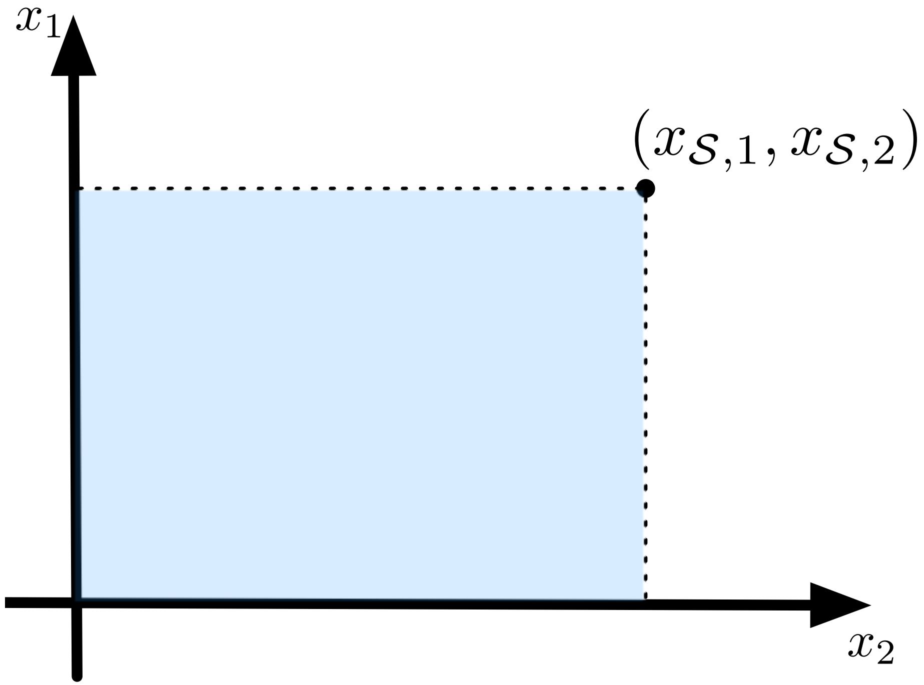


- By construction:
  1. $\tilde{\Psi}_{cmh}(x_S)$ is monotonically decreasing
  2. $\tilde{\Psi}_{cmh}(x_S)$ is a lower bound of $\Psi_{cmh}(x_S)$

Theorem 1: If $\tilde{\Psi}_{cmh}(x_S) > \delta$, then all superset feature combinations $\mathcal{S}' \supseteq \mathcal{S}$ are untestable and can be pruned from the search space

The lower envelope $\widetilde{\Psi}_{cmh}(x_S)$ for the CMH test can be evaluated in $O(k \log k)$ time

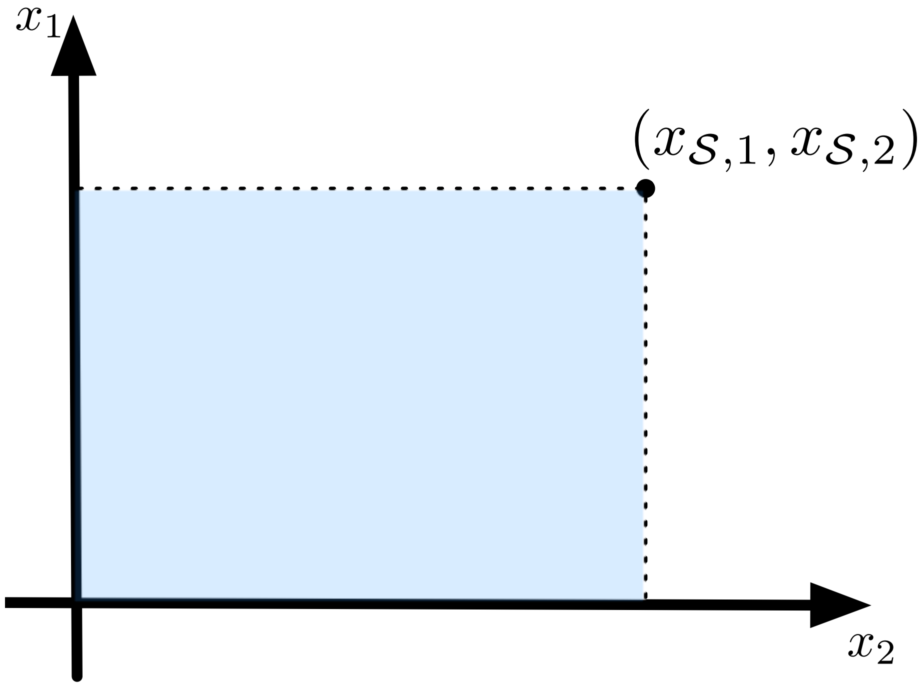
The lower envelope $\widetilde{\Psi}_{cmh}(x_S)$ for the CMH test can be evaluated in $O(k \log k)$ time



$$\tilde{\Psi}_{cmh}(\mathbf{x}_{\mathcal{S}}) = \min_{\mathbf{x}_{\mathcal{S}'} \leq \mathbf{x}_{\mathcal{S}}} \Psi_{cmh}(\mathbf{x}_{\mathcal{S}'})$$

The lower envelope $\widetilde{\Psi}_{cmh}(x_{\mathcal{S}})$ for the CMH test can be evaluated in $O(k \log k)$ time

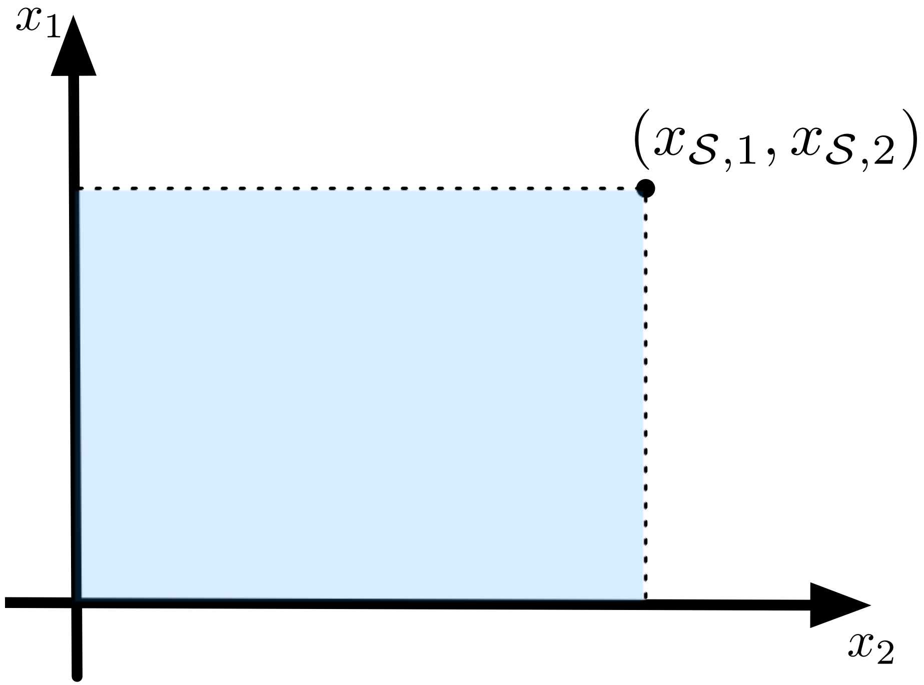
- Naively, computing $\widetilde{\Psi}_{cmh}(x_{\mathcal{S}})$ would require $O(m^k)$ runtime



$$\tilde{\Psi}_{cmh}(\mathbf{x}_{\mathcal{S}}) = \min_{\mathbf{x}_{\mathcal{S}'} \leq \mathbf{x}_{\mathcal{S}}} \Psi_{cmh}(\mathbf{x}_{\mathcal{S}'})$$

The lower envelope $\widetilde{\Psi}_{cmh}(\boldsymbol{x}_{\mathcal{S}})$ for the CMH test can be evaluated in $\boldsymbol{O}(\boldsymbol{k}\log\boldsymbol{k})$ time

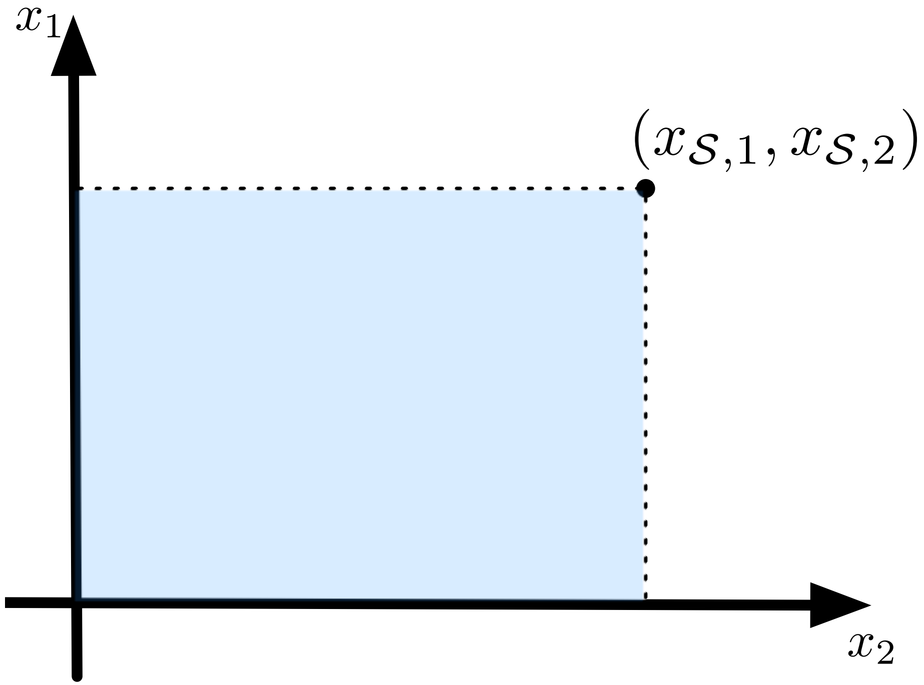
- Naively, computing $\widetilde{\Psi}_{cmh}(\boldsymbol{x}_{\mathcal{S}})$ would require $\boldsymbol{O}(\boldsymbol{m}^{\boldsymbol{k}})$ runtime
- $\widetilde{\Psi}_{cmh}(\boldsymbol{x}_{\mathcal{S}})$ must be evaluated once for each enumerated feature combination



$$\tilde{\Psi}_{cmh}(\mathbf{x}_{\mathcal{S}}) = \min_{\mathbf{x}_{\mathcal{S}'} \leq \mathbf{x}_{\mathcal{S}}} \Psi_{cmh}(\mathbf{x}_{\mathcal{S}'})$$

The lower envelope $\widetilde{\Psi}_{cmh}(x_S)$ for the CMH test can be evaluated in $O(k \log k)$ time

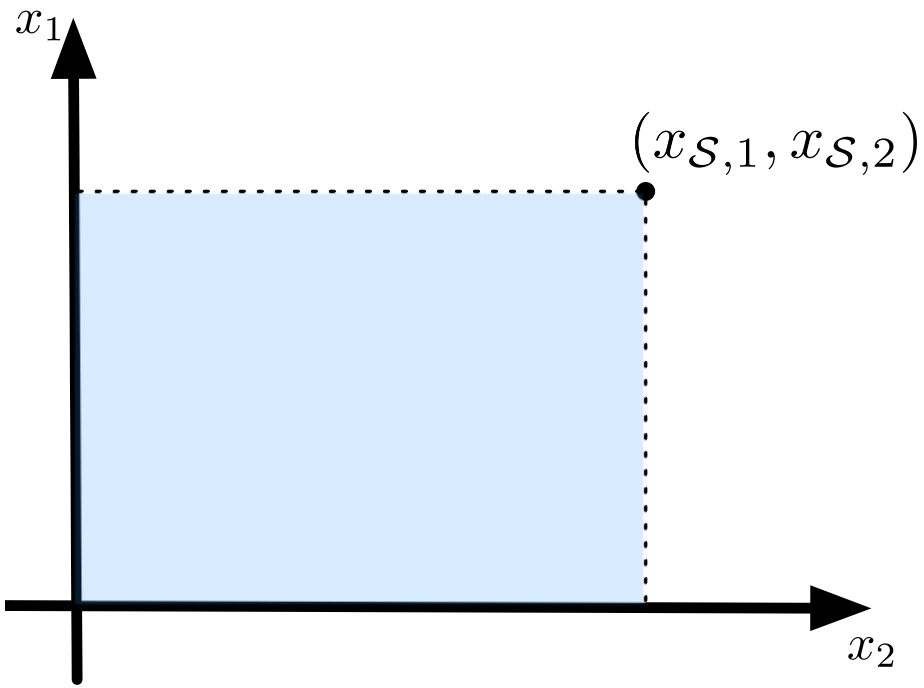
- Naively, computing $\widetilde{\Psi}_{cmh}(x_S)$ would require $O(m^k)$ runtime
- $\widetilde{\Psi}_{cmh}(x_S)$ must be evaluated once for each enumerated feature combination
- Need an efficient algorithm to compute $\widetilde{\Psi}_{cmh}(x_S)$



$$\tilde{\Psi}_{cmh}(\mathbf{x}_{\mathcal{S}}) = \min_{\mathbf{x}_{\mathcal{S}'} \leq \mathbf{x}_{\mathcal{S}}} \Psi_{cmh}(\mathbf{x}_{\mathcal{S}'})$$

The lower envelope $\widetilde{\Psi}_{cmh}(x_{\mathcal{S}})$ for the CMH test can be evaluated in $O(k \log k)$ time

- Naively, computing $\widetilde{\Psi}_{cmh}(x_{\mathcal{S}})$ would require $O(m^k)$ runtime
- $\widetilde{\Psi}_{cmh}(x_{\mathcal{S}})$ must be evaluated once for each enumerated feature combination
- Need an efficient algorithm to compute $\widetilde{\Psi}_{cmh}(x_{\mathcal{S}})$
- **Theorem 2**: $\widetilde{\Psi}_{cmh}(x_{\mathcal{S}})$ can be evaluated in $O(k \log k)$ time



$$\tilde{\Psi}_{cmh}(\mathbf{x}_{\mathcal{S}}) = \min_{\mathbf{x}_{\mathcal{S}'} \leq \mathbf{x}_{\mathcal{S}}} \Psi_{cmh}(\mathbf{x}_{\mathcal{S}'})$$

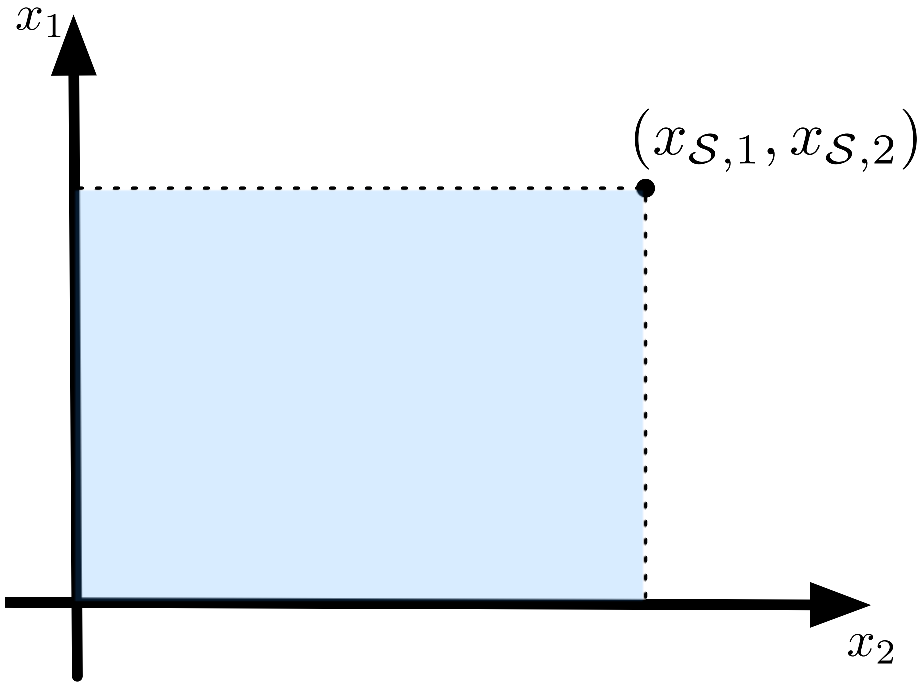
The lower envelope $\widetilde{\Psi}_{cmh}(\boldsymbol{x}_{\mathcal{S}})$ for the CMH test can be evaluated in $\boldsymbol{O}(\boldsymbol{k}\log\boldsymbol{k})$ time

- Naively, computing $\widetilde{\Psi}_{cmh}(\boldsymbol{x}_{\mathcal{S}})$ would require $\boldsymbol{O}(\boldsymbol{m}^{\boldsymbol{k}})$ runtime

- $\widetilde{\Psi}_{cmh}(\boldsymbol{x}_{\mathcal{S}})$ must be evaluated once for each enumerated feature combination

- Need an efficient algorithm to compute $\widetilde{\Psi}_{cmh}(\boldsymbol{x}_{\mathcal{S}})$

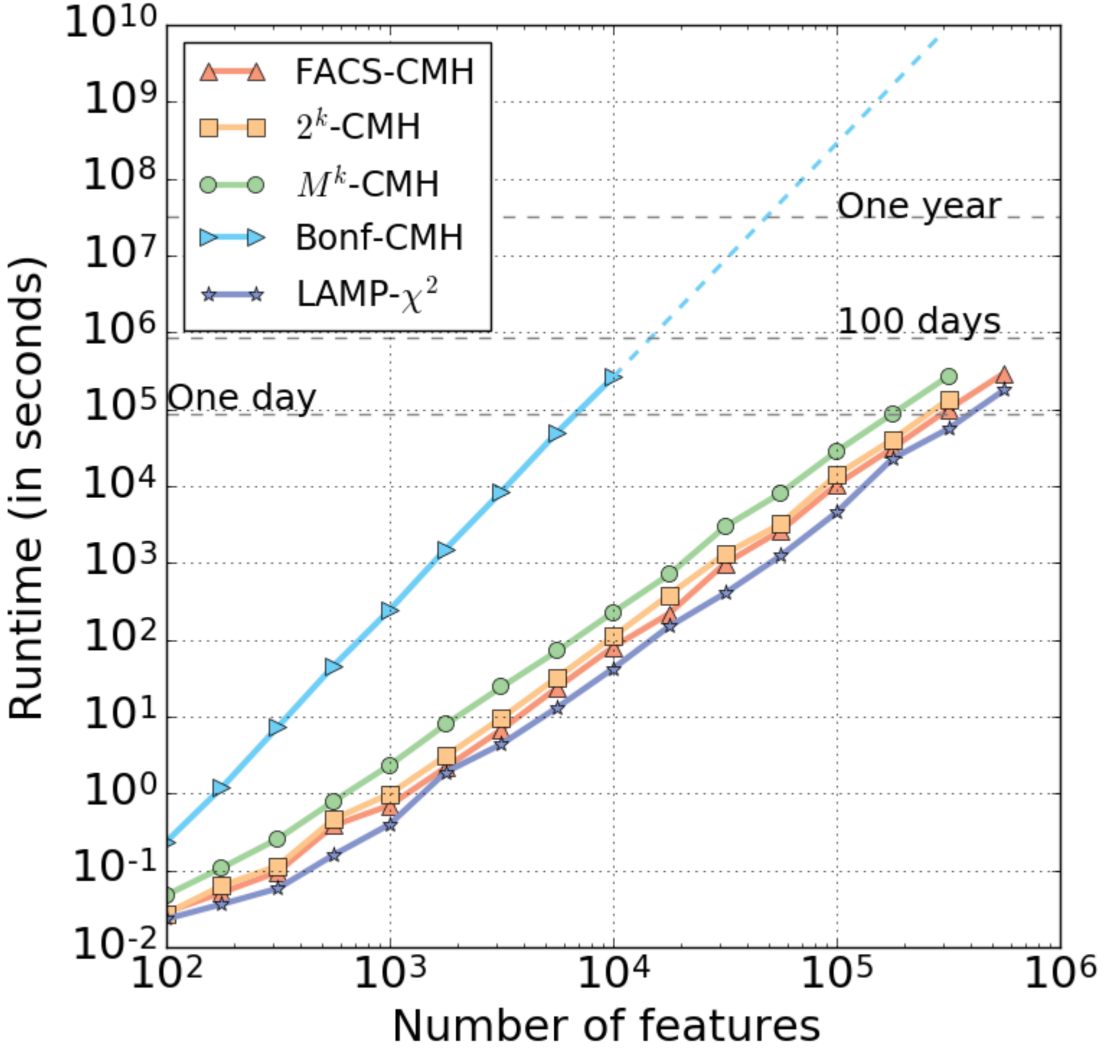
- **Theorem 2:** $\widetilde{\Psi}_{cmh}(\boldsymbol{x}_{\mathcal{S}})$ can be evaluated in $O(k \; \log k)$ time

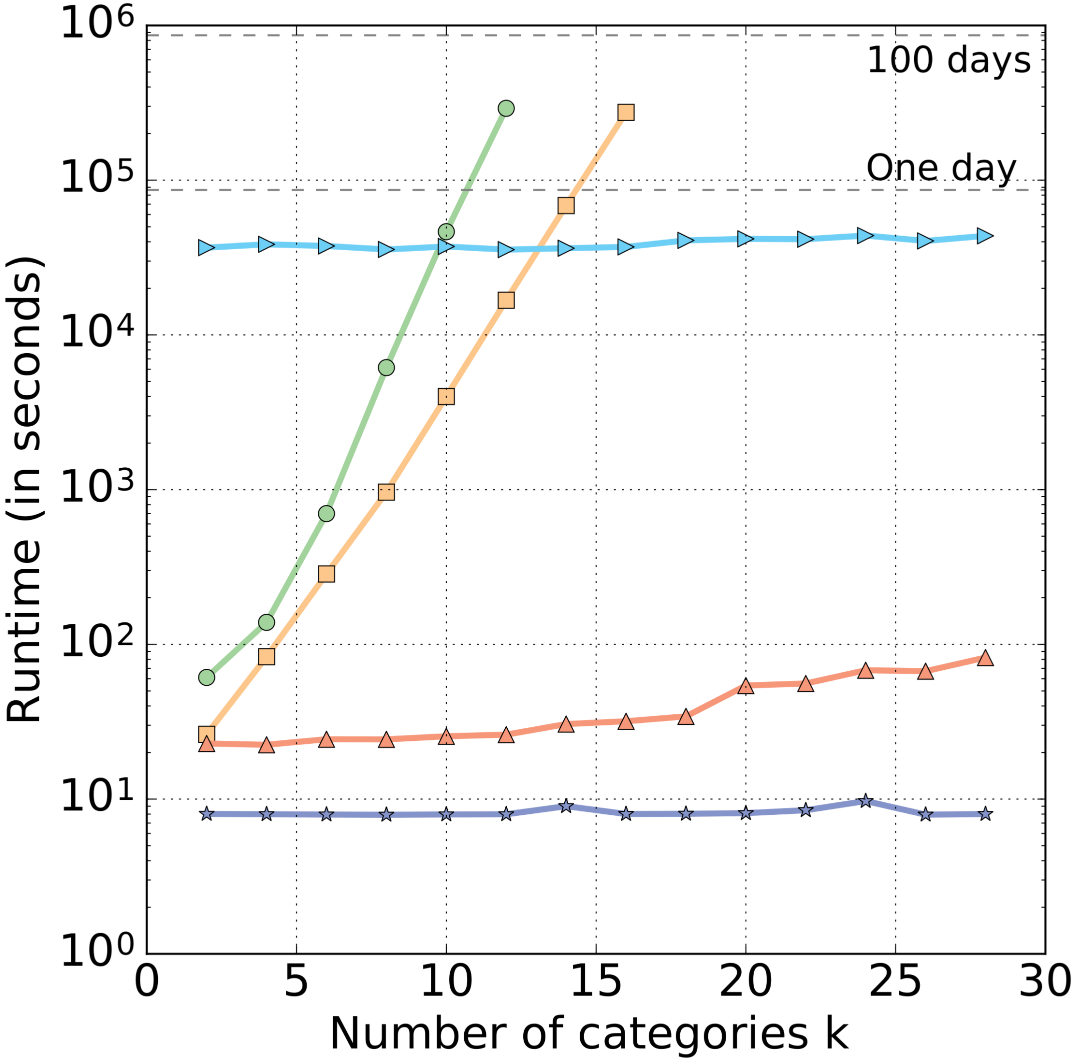


$$\tilde{\Psi}_{cmh}(\mathbf{x}_{\mathcal{S}}) = \min_{\mathbf{x}_{\mathcal{S}'} \leq \mathbf{x}_{\mathcal{S}}} \Psi_{cmh}(\mathbf{x}_{\mathcal{S}'})$$

Fast Automatic Conditional Search (FACS): An algorithm for significant pattern mining that can correct for a categorical covariate using the CMH test

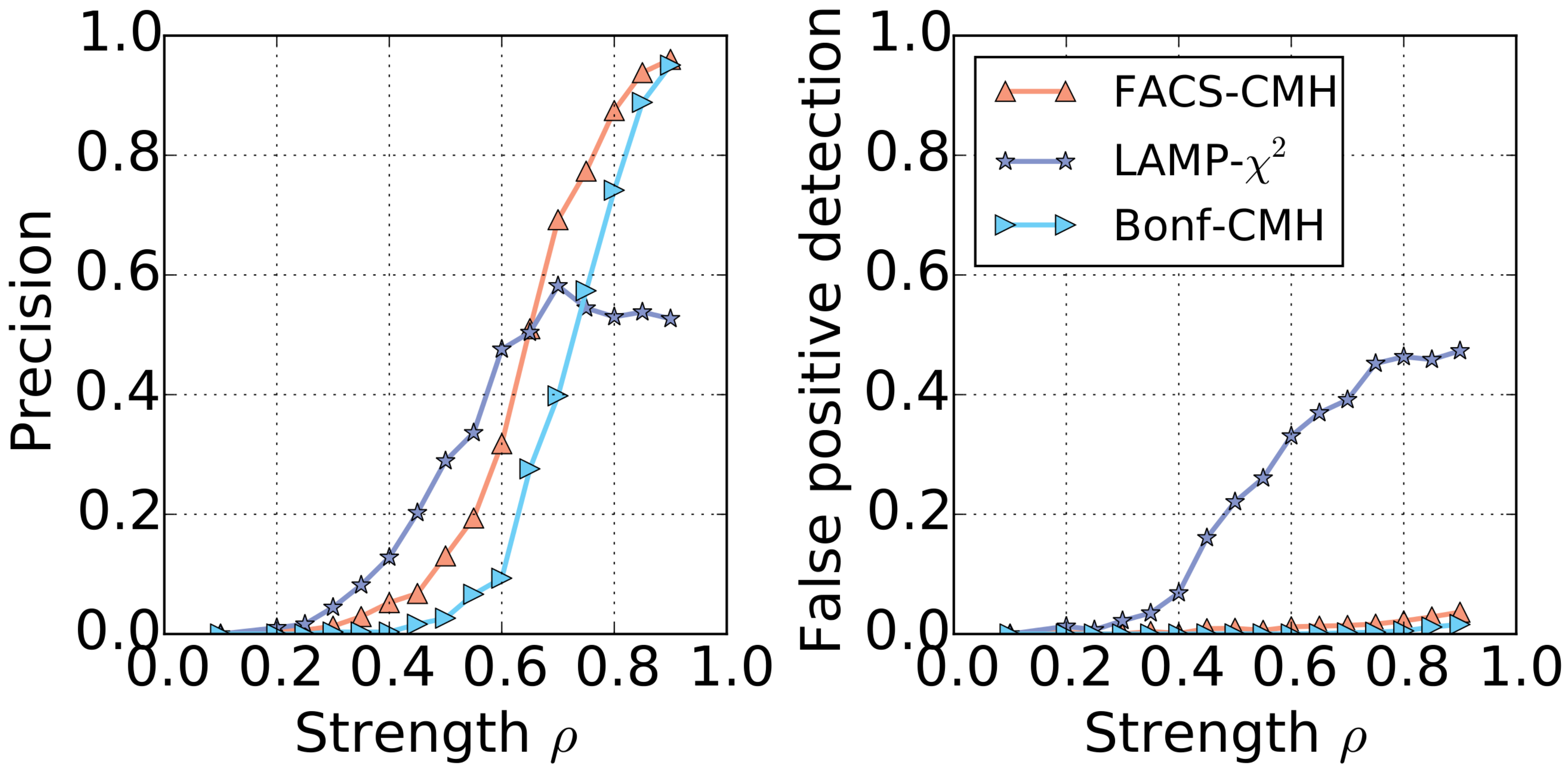
Correcting for covariates only leads to a negligible increase in runtime





ETH zürich

FACS successfully corrects for confounding without losing statistical power



Conclusions and outlook

- Significant pattern mining allows exploring all possible combinatorial feature interactions

- Significant pattern mining extends beyond simply (multiplicative) feature interactions
  - Significant subgraph mining (Sugiyama et al., SDM 2014)
  - Significant interval mining (Llinares-López et al., ISMB 2015)

- Significant pattern mining is a tool of great use for data exploration in personalized medicine

Conclusions and outlook

- Recent advances solve certain limitations of the first generation of significant pattern mining algorithms:
  - **Accounting for the dependence between feature interactions**: (Llinares-López et al., KDD 2015)
  - **Correcting for an observed categorical covariate**: (Papaxanthos et al., NIPS 2016)

- **Remaining challenges:**
  - Incorporating continuous data without discretization
  - Compression techniques to aid interpretability of the results

Acknowledgements

- **Karsten Borgwardt**

- **Laetitia Papaxanthos**
- **Mahito Sugiyama**
- **Dean Bodenham**

- **MLCB group:**
  - Damián Roqueiro
  - Dominik Grimm
  - Udo Gieraths
  - Anja Gumpinger
  - Lukas Folkman
  - Elisabetta Ghisu
  - Xiao He
  - Thomas Gumbsch
  - Caroline Weis
  - Katharina Heinrich

Presentation template by Dominik Grimm, icons from flaticon.com

Questions

Bibliography

References

[1] MacKay, David J. C. - **Information Theory, Inference and Learning Algorithms**. *Cambridge University Press*. 2002

[2] Tarone, R. E. – **A Modified Bonferroni Method for Discrete Data.** *Biometrics 46, 512-522*. 1990

[3] Terada, Aika et al. – **Statistical significance of combinatorial regulations.** *PNAS.* 2013

[4] Terada, Aika et al. – **Fast Westfall-Young Permutation Procedure for Combinatorial Regulation Discovery**. *IEEE International Conference on Bioinformatics and Biomedicine*. 2013

[5] Sugiyama, Mahito et al. - **Mining significant subgraphs with multiple testing correction**. *SIAM SDM*. 2015

[6] Llinares-López et al. - **Genome-wide detection of intervals of genetic heterogeneity associated with complex traits**. *ISMB*. 2015

[7] Llinares-López et al. – **Fast and Memory-Efficient Significant Pattern Mining via Permutation Testing**. *KDD*. 2015

[8] Papaxanthos et al. – **Finding significant combinations of features in the presence of categorical covariates**. *NIPS*. 2016

Appendix

The Algorithm (Westfall-Young Light)

- **Input:** Feature matrix $U \in \{0,1\}^{n \times p}$, class labels $y \in \{0,1\}^n$, target FWER α, number of permutations J
- **Initialization:**
  1. Compute and store J independent random permutations of the vector of class labels y
  2. $\delta \longleftarrow 1$
  3. $p_{min}^{(i)} \longleftarrow 1 \; \forall \, i = 1,2,\dots J$
- **DFS**$(\emptyset)$
- **Return** $\lfloor \alpha J \rfloor$ **smallest** $p_{min}^{(i)}$

- **DFS**$(\mathcal{S})$**:**
  1. $p_{min}^{(i)} \longleftarrow \min\left\{p_{min}^{(i)}, p^{(i)}(z_{\mathcal{S}})\right\} \forall \, i = 1,2,\dots J$ # Update minimum p-value so far for each permutation
  2. $FWER_{wy}(\delta) \longleftarrow \frac{1}{J}\sum_{i=1}^{J} \mathbf{1}\left[p_{min}^{(i)} \leq \delta\right]$ # Compute lower bound on FWER based on minimum p-values so far
  3. While $FWER_{wy}(\delta) > \alpha$: # If FWER condition is violated, decrease significance threshold until restored
     - Decrease δ
     - $FWER_{wy}(\delta) \longleftarrow \frac{1}{J}\sum_{i=1}^{J} \mathbf{1}\left[p_{min}^{(i)} \leq \delta\right]$
  4. For $\mathcal{S}' \in \mathrm{Children}(\mathcal{S})$: # Continue depth-first search recursively
     - If $\Psi(x_{\mathcal{S}'}) \leq \delta$: # Search-space pruning condition!
       - DFS$(\mathcal{S}')$