CS294-248 Special Topics in Database Theory
Unit 5 (Part 2): Database Constraints

Dan Suciu

University of Washington

Outline

- Classical constraints: FDs, MVDs, CIs

- The basics, and a modern approach

Functional Dependencies

Functional Dependencies

Fix a relation schema $R(\boldsymbol{X})$.

A Functional Dependency, FD, is an expression $\boldsymbol{U} \rightarrow \boldsymbol{V}$ for $\boldsymbol{U}, \boldsymbol{V} \subseteq \boldsymbol{X}$.

We say that an instance R^D satisfies the FD σ, and write $R^D \models \sigma$, if:

$$\forall t, t' \in R^D : \quad t.\boldsymbol{U} = t'\boldsymbol{U} \Rightarrow t.\boldsymbol{V} = t'.\boldsymbol{V}$$

If Σ is a set of FDs, then we write $R^D \models \Sigma$ if $R^D \models \sigma$ for all $\sigma \in \Sigma$.

Example

X	Y	Z
123	12	23
321	32	21
125	12	25
323	32	23
637	63	37
283	28	83

Then:

$$
\begin{aligned}
R^D \models & X \rightarrow Y, \\
& X \rightarrow Z, \\
& X \rightarrow YZ, \\
& YZ \rightarrow X
\end{aligned}
$$

But:

$$R^D \not\models Y \rightarrow X$$

The Implication Problem

We say that a set of FDs Σ implies and FD σ if $\forall R^D$, $R^D \models \Sigma$ implies $R^D \models \sigma$.

$$\boxed{\Sigma \models \sigma}$$

The Implication Problem

We say that a set of FDs Σ implies and FD σ if $\forall R^D$, $R^D \models \Sigma$ implies $R^D \models \sigma$.

$$\boxed{\Sigma \models \sigma}$$

Example: $AB \to C, CD \to E \models ABD \to E$.

The Implication Problem

We say that a set of FDs Σ implies and FD σ if $\forall R^D$, $R^D \models \Sigma$ implies $R^D \models \sigma$.

$$\boxed{\Sigma \models \sigma}$$

Example: $AB \to C, CD \to E \models ABD \to E$.

Proof:

A	B	C	D	E
		$\cdots$		
a	b	y	d	?
		$\cdots$		
a	b	v	d	?
		$\cdots$		

Armstrong's Axioms

Many minor variations. My favorite:

Trivial: $\models \boldsymbol{UV} \rightarrow \boldsymbol{U}$
Transitivity: $\boldsymbol{U} \rightarrow \boldsymbol{V}, \boldsymbol{V} \rightarrow \boldsymbol{W} \models \boldsymbol{U} \rightarrow \boldsymbol{W}$
Splitting/combining: $\boldsymbol{U} \rightarrow \boldsymbol{VW}$ iff $\boldsymbol{U} \rightarrow \boldsymbol{V}, \boldsymbol{U} \rightarrow \boldsymbol{W}$

However, cumbersome to use: Can we check $\Sigma \models \sigma$ in PTIME?

The Closure Operator

Fix Σ. The closure of a set $\boldsymbol{U}$ is $\boxed{\boldsymbol{U}^+ \stackrel{\text{def}}{=} \{Z \mid \Sigma \models \boldsymbol{U} \rightarrow Z\}}$

Note that Σ is implicit in defining $\boldsymbol{U}^+$.

Databases 101 (to discuss in class):

- Given $\boldsymbol{U}$, one can compute the closure $\boldsymbol{U}^+$ in PTIME
- $\Sigma \models \boldsymbol{U} \rightarrow \boldsymbol{V}$ iff $\boldsymbol{V} \subseteq \boldsymbol{U}^+$.
- Example: $\Sigma = \{AB \rightarrow C, CD \rightarrow E\}$;
  $AD^+ =?$

The Closure Operator

Fix Σ. The closure of a set $\boldsymbol{U}$ is $\boxed{\boldsymbol{U}^+ \stackrel{\text{def}}{=} \{Z \mid \Sigma \models \boldsymbol{U} \to Z\}}$

Note that Σ is implicit in defining $\boldsymbol{U}^+$.

Databases 101 (to discuss in class):

- Given $\boldsymbol{U}$, one can compute the closure $\boldsymbol{U}^+$ in PTIME
- $\Sigma \models \boldsymbol{U} \to \boldsymbol{V}$ iff $\boldsymbol{V} \subseteq \boldsymbol{U}^+$.
- Example: $\Sigma = \{AB \to C, CD \to E\}$;
  $AD^+ =?$ AD

The Closure Operator

Fix Σ. The closure of a set $\boldsymbol{U}$ is $\boxed{\boldsymbol{U}^+ \stackrel{\text{def}}{=} \{Z \mid \Sigma \models \boldsymbol{U} \to Z\}}$

Note that Σ is implicit in defining $\boldsymbol{U}^+$.

Databases 101 (to discuss in class):

- Given $\boldsymbol{U}$, one can compute the closure $\boldsymbol{U}^+$ in PTIME
- $\Sigma \models \boldsymbol{U} \to \boldsymbol{V}$ iff $\boldsymbol{V} \subseteq \boldsymbol{U}^+$.
- Example: $\Sigma = \{AB \to C, CD \to E\}$;
  $AD^+ =?\ AD$ $ABD^+ =?$

The Closure Operator

Fix Σ. The closure of a set $\boldsymbol{U}$ is $\boxed{\boldsymbol{U}^+ \stackrel{\text{def}}{=} \{Z \mid \Sigma \models \boldsymbol{U} \rightarrow Z\}}$

Note that Σ is implicit in defining $\boldsymbol{U}^+$.

Databases 101 (to discuss in class):

- Given $\boldsymbol{U}$, one can compute the closure $\boldsymbol{U}^+$ in PTIME
- $\Sigma \models \boldsymbol{U} \rightarrow \boldsymbol{V}$ iff $\boldsymbol{V} \subseteq \boldsymbol{U}^+$.
- Example: $\Sigma = \{AB \rightarrow C, CD \rightarrow E\}$;
  $AD^+ =?$ AD $ABD^+ =?$ $ABCD$.

2-Tuple Relation

Fact

If $\Sigma \not\models \sigma$ then there exists a 2-tuple relation R s.t. $R \models \Sigma$ and $R \not\models \sigma$.

Example: $AB \rightarrow C, CD \rightarrow E \not\models CD \rightarrow A$.

Find a counterexample with 2 tuples (use values $0, 1$):

$R =$

A	B	C	D	E
?	?	?	?	?
?	?	?	?	?

2-Tuple Relation

Fact

If $\Sigma \not\models \sigma$ then there exists a 2-tuple relation R s.t. $R \models \Sigma$ and $R \not\models \sigma$.

Example: $AB \to C, CD \to E \not\models CD \to A$.

Find a counterexample with 2 tuples (use values $0, 1$): $CD^+ = CDE$

$R =$

A	B	C	D	E
0	0	0	0	0
1	1	0	0	0

2-Tuple Relation

Fact

If $\Sigma \not\models \sigma$ then there exists a 2-tuple relation R s.t. $R \models \Sigma$ and $R \not\models \sigma$.

Example: $AB \rightarrow C, CD \rightarrow E \not\models CD \rightarrow A$.

Find a counterexample with 2 tuples (use values $0, 1$): $CD^+ = CDE$

$R =$

A	B	C	D	E
0	0	0	0	0
1	1	0	0	0

To refute $\boldsymbol{U} \rightarrow \boldsymbol{V}$: Tuple 1: $(0, 0, \ldots, 0)$, Tuple 2: $\boldsymbol{U}^+ := 0$, rest $:= 1$.

Armstrong Relation

- We can refute a single implication $\Sigma \models \sigma$ using a 2-tuple relation.
- Armstrong relation for Σ is a relation R_Σ that refutes *all* FDs not implied by Σ.
- Equivalently, $\Sigma \models \sigma$ iff $R_\Sigma \models \sigma$.
- The construction of R_Σ is more interesting that the application. Next.

The Direct Product

[Fagin, 1982]
The direct product[1] of two tuples $t = (a_1, \ldots, a_n)$ and $t' = (b_1, \ldots, b_n)$ is:

$$t \otimes t' \stackrel{\text{def}}{=} ((a_1, b_1), \ldots, (a_n, b_n))$$

[1] A.k.a. domain product.

The Direct Product

[Fagin, 1982]
The direct product[1] of two tuples $t = (a_1, \ldots, a_n)$ and $t' = (b_1, \ldots, b_n)$ is:

$$t \otimes t' \stackrel{\text{def}}{=} ((a_1, b_1), \ldots, (a_n, b_n))$$

Notice: the domain of $t \otimes t'$ is the cartesian product of domains of t and t'.

[1] A.k.a. domain product.

The Direct Product

[Fagin, 1982]
The direct product[1] of two tuples $t = (a_1, \ldots, a_n)$ and $t' = (b_1, \ldots, b_n)$ is:

$$t \otimes t' \stackrel{\text{def}}{=} ((a_1, b_1), \ldots, (a_n, b_n))$$

Notice: the domain of $t \otimes t'$ is the cartesian product of domains of t and t'.

The direct product of two relations $R(\boldsymbol{X}), R'(\boldsymbol{X})$ (same attributes!) is

$$R \otimes R' \stackrel{\text{def}}{=} \{t \otimes t' \mid t \in R, t' \in R'\}$$

[1] A.k.a. domain product.

Example: Cartesian Product v.s. Direct Product

$T =$

A	B
1	5
1	6

$S =$

X	Y	Z
a	b	c
f	b	d
a	e	d

Example: Cartesian Product v.s. Direct Product

$T =$

A	B
1	5
1	6

$S =$

X	Y	Z
a	b	c
f	b	d
a	e	d

$T \times S =$

A	B	X	Y	Z
1	5	a	b	c
1	6	a	b	c
1	5	f	b	d
		$\ldots$		

Example: Cartesian Product v.s. Direct Product

$T =$

A	B
1	5
1	6

$S =$

X	Y	Z
a	b	c
f	b	d
a	e	d

$T \times S =$

A	B	X	Y	Z
1	5	a	b	c
1	6	a	b	c
1	5	f	b	d
		$\ldots$		

$R =$

X	Y	Z
1	5	m
1	6	m

Example: Cartesian Product v.s. Direct Product

$T =$

A	B
1	5
1	6

$S =$

X	Y	Z
a	b	c
f	b	d
a	e	d

$T \times S =$

A	B	X	Y	Z
1	5	a	b	c
1	6	a	b	c
1	5	f	b	d
...				

$R =$

X	Y	Z
1	5	m
1	6	m

$R \otimes S =$

X	Y	Z
$1a$	$5b$	mc
$1a$	$6b$	mc
$2a$	$6b$	nc
$1f$	$5b$	md
...		

Example: Cartesian Product v.s. Direct Product

$T =$

A	B
1	5
1	6

$S =$

X	Y	Z
a	b	c
f	b	d
a	e	d

$T \times S =$

A	B	X	Y	Z
1	5	a	b	c
1	6	a	b	c
1	5	f	b	d
...				

$R =$

X	Y	Z
1	5	m
1	6	m

$R \otimes S =$

X	Y	Z
$1a$	$5b$	mc
$1a$	$6b$	mc
$2a$	$6b$	nc
$1f$	$5b$	md
...		

Given prob. distributions with entropies h_R, h_S, what is $h_{R \otimes S}$?
In class.

Example: Cartesian Product v.s. Direct Product

$T =$

A	B
1	5
1	6

$S =$

X	Y	Z
a	b	c
f	b	d
a	e	d

$T \times S =$

A	B	X	Y	Z
1	5	a	b	c
1	6	a	b	c
1	5	f	b	d
...				

$R =$

X	Y	Z
1	5	m
1	6	m

$R \otimes S =$

X	Y	Z
$1a$	$5b$	mc
$1a$	$6b$	mc
$2a$	$6b$	nc
$1f$	$5b$	md
...		

Given prob. distributions with entropies h_R, h_S, what is $h_{R \otimes S}$?
In class. $h_R + h_S$ (sum of two vectors).
h_T, h_S cannot be added, since they have $2^2, 2^3$ dimensions.

Armstrong's Relation

Lemma

For any FD σ, $R \otimes R' \models \sigma$ iff $R \models \sigma$ and $R' \models \sigma$.

Proof in class (it's straightforward).

Armstrong's Relation

Lemma

For any FD σ, $R \otimes R' \models \sigma$ iff $R \models \sigma$ and $R' \models \sigma$.

Proof in class (it's straightforward).

Theorem (Armstrong's Relation)

For any set of FDs Σ there exists R_Σ s.t., for any FD σ, $\Sigma \models \sigma$ iff $R_\Sigma \models \sigma$.

Armstrong's Relation

Lemma

For any FD σ, $R \otimes R' \models \sigma$ iff $R \models \sigma$ and $R' \models \sigma$.

Proof in class (it's straightforward).

Theorem (Armstrong's Relation)

For any set of FDs Σ there exists R_Σ s.t., for any FD σ, $\Sigma \models \sigma$ iff $R_\Sigma \models \sigma$.

Proof Let $\sigma_i, i = 1, n$ be all FDs not implied by Σ.

Since $\Sigma \not\models \sigma_i$, there exists a 2-tuple R_i such that $R_i \models \Sigma$ and $R_i \not\models \sigma$.

Armstrong's Relation

Lemma

For any FD σ, $R \otimes R' \models \sigma$ iff $R \models \sigma$ and $R' \models \sigma$.

Proof in class (it's straightforward).

Theorem (Armstrong's Relation)

For any set of FDs Σ there exists R_Σ s.t., for any FD σ, $\Sigma \models \sigma$ iff $R_\Sigma \models \sigma$.

Proof Let $\sigma_i, i = 1, n$ be all FDs not implied by Σ.

Since $\Sigma \not\models \sigma_i$, there exists a 2-tuple R_i such that $R_i \models \Sigma$ and $R_i \not\models \sigma$.

Then $R_\Sigma \stackrel{\text{def}}{=} R_1 \otimes \cdots \otimes R_n$ satisfies the theorem. Why?

Armstrong's Relation

Lemma

For any FD σ, $R \otimes R' \models \sigma$ iff $R \models \sigma$ and $R' \models \sigma$.

Proof in class (it's straightforward).

Theorem (Armstrong's Relation)

For any set of FDs Σ there exists R_Σ s.t., for any FD σ, $\Sigma \models \sigma$ iff $R_\Sigma \models \sigma$.

Proof Let $\sigma_i, i = 1, n$ be all FDs not implied by Σ.

Since $\Sigma \not\models \sigma_i$, there exists a 2-tuple R_i such that $R_i \models \Sigma$ and $R_i \not\models \sigma$.

Then $R_\Sigma \stackrel{\text{def}}{=} R_1 \otimes \cdots \otimes R_n$ satisfies the theorem. Why?

How large is R_Σ?

Discussion

Next:

- Defining the FDs is equivalent to defining the closure operator $\boldsymbol{U}^+$.
- In turn, this is equivalent to defining the *closed* sets, i.e. those that satisfy $\boldsymbol{U} = \boldsymbol{U}^+$.
- And this is equivalent to defining the lattice of closed elements.

The Closure Operator: Properties

Monotone: If $\boldsymbol{U} \subseteq \boldsymbol{V}$, then $\boldsymbol{U}^+ \subseteq \boldsymbol{V}^+$. Why??

Expansive: $\boldsymbol{U} \subseteq \boldsymbol{U}^+$ Why??

Idempotent: $(\boldsymbol{U}^+)^+ = \boldsymbol{U}^+$ Why??

Wikipedia calls these properties increasing, extensive, idempotent.

Discussion

The closure operator, and its associated closure system occur in many areas of math and CS.

- For any subset $S \subseteq \mathbb{R}^d$, its linear span, $\texttt{span}(S)$, is the smallest vector space containing S; $\texttt{span}$ is a closure operator.
- For any subset $S \subseteq \mathbb{R}^d$, let $\texttt{convex}(S) \subseteq \mathbb{R}^d$ be its convex closure; $\texttt{convex}$ is a closure operator.
- The topological closure of a subset $S \subseteq \mathbb{R}^d$ is the set $\bar{S}$ consisting of all limits $\lim_n x_n$, where the sequence x_n is in S.
- Fix an algebra A. The algebra generated by a subset S is the smallest sub-algebra containing S.

Detour: Closure Operators

Fix a set Ω.

Detour: Closure Operators

Fix a set Ω.

Definition (Closure Operator)

A closure operator is $cl : \mathcal{P}(\Omega) \to \mathcal{P}(\Omega)$ that is:

- monotone $A \subseteq B \Rightarrow cl(A) \subseteq cl(B)$
- expansive $A \subseteq cl(A)$
- idempotent $cl(cl(A)) = cl(A)$

Detour: Closure Operators

Fix a set Ω.

Definition (Closure Operator)

A closure operator is $cl : \mathcal{P}(\Omega) \to \mathcal{P}(\Omega)$ that is:

- monotone $A \subseteq B \Rightarrow cl(A) \subseteq cl(B)$
- expansive $A \subseteq cl(A)$
- idempotent $cl(cl(A)) = cl(A)$

Definition (Closure System)

A closure system is $\mathcal{C} \subseteq \mathcal{P}(\Omega)$ s.t.

- for any $\mathcal{S} \subseteq \mathcal{C}$, $\bigcap \mathcal{S} \in \mathcal{C}$.

Detour: Closure Operators

Fix a set Ω.

Definition (Closure Operator)

A closure operator is $cl : \mathcal{P}(\Omega) \to \mathcal{P}(\Omega)$ that is:

- monotone $A \subseteq B \Rightarrow cl(A) \subseteq cl(B)$
- expansive $A \subseteq cl(A)$
- idempotent $cl(cl(A)) = cl(A)$

Definition (Closure System)

A closure system is $\mathcal{C} \subseteq \mathcal{P}(\Omega)$ s.t.

- for any $\mathcal{S} \subseteq \mathcal{C}$, $\bigcap \mathcal{S} \in \mathcal{C}$.

Equivalence

- Given $\mathcal{C}$, $cl(A) \stackrel{\text{def}}{=} \bigcap\{X \in \mathcal{C} \mid A \subseteq X\}$ is a closure operator.

Detour: Closure Operators

Fix a set Ω.

Definition (Closure Operator)

A closure operator is $cl : \mathcal{P}(\Omega) \to \mathcal{P}(\Omega)$ that is:

- monotone $A \subseteq B \Rightarrow cl(A) \subseteq cl(B)$
- expansive $A \subseteq cl(A)$
- idempotent $cl(cl(A)) = cl(A)$

Definition (Closure System)

A closure system is $\mathcal{C} \subseteq \mathcal{P}(\Omega)$ s.t.

- for any $\mathcal{S} \subseteq \mathcal{C}$, $\bigcap \mathcal{S} \in \mathcal{C}$.

Equivalence

- Given $\mathcal{C}$, $cl(A) \stackrel{\text{def}}{=} \bigcap \{X \in \mathcal{C} \mid A \subseteq X\}$ is a closure operator.
- Given cl, $\mathcal{C} \stackrel{\text{def}}{=} \{X \mid cl(X) = X\}$ is a closure system.

Detour: Closure Operators

Fix a set Ω.

Definition (Closure Operator)

A closure operator is $cl : \mathcal{P}(\Omega) \rightarrow \mathcal{P}(\Omega)$ that is:

- monotone $A \subseteq B \Rightarrow cl(A) \subseteq cl(B)$
- expansive $A \subseteq cl(A)$
- idempotent $cl(cl(A)) = cl(A)$

Definition (Closure System)

A closure system is $\mathcal{C} \subseteq \mathcal{P}(\Omega)$ s.t.

- for any $\mathcal{S} \subseteq \mathcal{C}$, $\bigcap \mathcal{S} \in \mathcal{C}$.

Equivalence

- Given $\mathcal{C}$, $cl(A) \stackrel{\text{def}}{=} \bigcap\{X \in \mathcal{C} \mid A \subseteq X\}$ is a closure operator.
- Given cl, $\mathcal{C} \stackrel{\text{def}}{=} \{X \mid cl(X) = X\}$ is a closure system.
  Proof: We check that $A \stackrel{\text{def}}{=} \bigcap \mathcal{S}$ is in $\mathcal{C}$, for any set $\mathcal{S} \subseteq \mathcal{C}$:

Detour: Closure Operators

Fix a set Ω.

Definition (Closure Operator)

A closure operator is $cl : \mathcal{P}(\Omega) \to \mathcal{P}(\Omega)$ that is:

- monotone $A \subseteq B \Rightarrow cl(A) \subseteq cl(B)$
- expansive $A \subseteq cl(A)$
- idempotent $cl(cl(A)) = cl(A)$

Definition (Closure System)

A closure system is $\mathcal{C} \subseteq \mathcal{P}(\Omega)$ s.t.

- for any $\mathcal{S} \subseteq \mathcal{C}$, $\bigcap \mathcal{S} \in \mathcal{C}$.

Equivalence

- Given $\mathcal{C}$, $cl(A) \stackrel{\text{def}}{=} \bigcap\{X \in \mathcal{C} \mid A \subseteq X\}$ is a closure operator.
- Given cl, $\mathcal{C} \stackrel{\text{def}}{=} \{X \mid cl(X) = X\}$ is a closure system.
  Proof: We check that $A \stackrel{\text{def}}{=} \bigcap \mathcal{S}$ is in $\mathcal{C}$, for any set $\mathcal{S} \subseteq \mathcal{C}$:
  $cl(A) = cl\left(\bigcap\{X \mid X \in \mathcal{S}\}\right) \subseteq cl(X)$ for all $X \in \mathcal{S}$.

Detour: Closure Operators

Fix a set Ω.

Definition (Closure Operator)

A closure operator is $cl : \mathcal{P}(\Omega) \rightarrow \mathcal{P}(\Omega)$ that is:

- monotone $A \subseteq B \Rightarrow cl(A) \subseteq cl(B)$
- expansive $A \subseteq cl(A)$
- idempotent $cl(cl(A)) = cl(A)$

Definition (Closure System)

A closure system is $\mathcal{C} \subseteq \mathcal{P}(\Omega)$ s.t.

- for any $\mathcal{S} \subseteq \mathcal{C}$, $\bigcap \mathcal{S} \in \mathcal{C}$.

Equivalence

- Given $\mathcal{C}$, $cl(A) \stackrel{\text{def}}{=} \bigcap \{X \in \mathcal{C} \mid A \subseteq X\}$ is a closure operator.
- Given cl, $\mathcal{C} \stackrel{\text{def}}{=} \{X \mid cl(X) = X\}$ is a closure system.

  Proof: We check that $A \stackrel{\text{def}}{=} \bigcap \mathcal{S}$ is in $\mathcal{C}$, for any set $\mathcal{S} \subseteq \mathcal{C}$:
  $cl(A) = cl\left(\bigcap\{X \mid X \in \mathcal{S}\}\right) \subseteq cl(X)$ for all $X \in \mathcal{S}$.
  Therefore $cl(A) \subseteq \bigcap\{X \mid X \in \mathcal{S}\} = A$.

From FDs to the Lattice of Closed Sets

A set of FDs for $R(\boldsymbol{X})$ is equivalent to as closure system on $\boldsymbol{X}$.

Moreover, a closure system $\mathcal{C}$ forms a lattice, $(\mathcal{C}, \wedge, \vee)$:

$$X \wedge Y \stackrel{\text{def}}{=} X \cap Y \qquad X \vee Y \stackrel{\text{def}}{=} (X \cup Y)^+$$

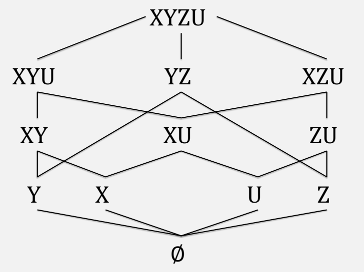
From FDs to the Lattice of Closed Sets

A set of FDs for $R(\boldsymbol{X})$ is equivalent to as closure system on $\boldsymbol{X}$.

Moreover, a closure system $\mathcal{C}$ forms a lattice, $(\mathcal{C}, \wedge, \vee)$:

$$X \wedge Y \stackrel{\text{def}}{=} X \cap Y \qquad\qquad X \vee Y \stackrel{\text{def}}{=} (X \cup Y)^+$$

Example: $YU \to X, XZ \to U$



Discussion

- Functional dependencies are a key concept in CS, beyond databases.

- In databases, the have two traditional applications:
  - Database normalization: BCNF, 3NF
  - Keys/foreign keys; "semantic pointers"

- More recent applications: discover FDs from data, approximate FDs, repairing for FDs (data imputation).

Multivalued Dependencies

Relation Decomposition

Take a relation R, partition its variables into $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}$.

Instead of storing $R(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W})$ we store its projections:

$$R_1(\boldsymbol{U}, \boldsymbol{V}) \stackrel{\text{def}}{=} \Pi_{\boldsymbol{UV}}(R), \; R_2(\boldsymbol{U}, \boldsymbol{W}) \stackrel{\text{def}}{=} \Pi_{\boldsymbol{UW}}(R)$$

Can we always recover R from $R_1 \bowtie R_2$?

Relation Decomposition

Take a relation R, partition its variables into $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}$.

Instead of storing $R(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W})$ we store its projections:

$$R_1(\boldsymbol{U}, \boldsymbol{V}) \stackrel{\text{def}}{=} \Pi_{\boldsymbol{UV}}(R),\ R_2(\boldsymbol{U}, \boldsymbol{W}) \stackrel{\text{def}}{=} \Pi_{\boldsymbol{UW}}(R)$$

Can we always recover R from $R_1 \bowtie R_2$? NO! In general $R \subseteq R_1 \bowtie R_2$.

Relation Decomposition

Take a relation R, partition its variables into $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}$.

Instead of storing $R(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W})$ we store its projections:

$$R_1(\boldsymbol{U}, \boldsymbol{V}) \stackrel{\text{def}}{=} \Pi_{\boldsymbol{UV}}(R),\; R_2(\boldsymbol{U}, \boldsymbol{W}) \stackrel{\text{def}}{=} \Pi_{\boldsymbol{UW}}(R)$$

Can we always recover R from $R_1 \bowtie R_2$? NO! In general $R \subseteq R_1 \bowtie R_2$.

Lossless decomposition: when $R = R_1 \bowtie R_2$.

Relation Decomposition

Take a relation R, partition its variables into $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W}$.

Instead of storing $R(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W})$ we store its projections:

$$R_1(\boldsymbol{U}, \boldsymbol{V}) \stackrel{\text{def}}{=} \Pi_{\boldsymbol{UV}}(R),\ R_2(\boldsymbol{U}, \boldsymbol{W}) \stackrel{\text{def}}{=} \Pi_{\boldsymbol{UW}}(R)$$

Can we always recover R from $R_1 \bowtie R_2$? NO! In general $R \subseteq R_1 \bowtie R_2$.

Lossless decomposition: when $R = R_1 \bowtie R_2$.

Fact If $\boldsymbol{U} \rightarrow \boldsymbol{V}$ holds then the decomposition is lossless. This is the basis of *database normalization* (BCNF, 3NF).

Multivalued Dependency

A multivalued dependency is $\boldsymbol{U} \twoheadrightarrow \boldsymbol{V}$.

Multivalued Dependency

A multivalued dependency is $\boldsymbol{U} \twoheadrightarrow \boldsymbol{V}$.

A relation $R(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W})$ satisfies the MVD, if:

$$R = \Pi_{\boldsymbol{UV}}(R) \bowtie \Pi_{\boldsymbol{UW}}(R)$$

Multivalued Dependency

A multivalued dependency is $\boldsymbol{U} \twoheadrightarrow \boldsymbol{V}$.

A relation $R(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W})$ satisfies the MVD, if:

$$R = \Pi_{\boldsymbol{UV}}(R) \bowtie \Pi_{\boldsymbol{UW}}(R)$$

We will always denote the MVD by $\boldsymbol{U} \twoheadrightarrow \boldsymbol{V}; \boldsymbol{W}$ ($\boldsymbol{W} \stackrel{\text{def}}{=}$ the rest of attrs).

Multivalued Dependency

A multivalued dependency is $\boldsymbol{U} \twoheadrightarrow \boldsymbol{V}$.

A relation $R(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W})$ satisfies the MVD, if:

$$R = \Pi_{\boldsymbol{UV}}(R) \bowtie \Pi_{\boldsymbol{UW}}(R)$$

We will always denote the MVD by $\boldsymbol{U} \twoheadrightarrow \boldsymbol{V}; \boldsymbol{W}$ ($\boldsymbol{W} \stackrel{\text{def}}{=}$ the rest of attrs).

Equivalently: if $(u, v_1, w_2), (u, v_2, w_2) \in R$ then $(u, v_1, w_2) \in R$ (and, by symmetry, $(u, v_2, w_1) \in R$).

Examples

1. Fix $R(X, Y, Z)$. If $Z \to X$, then $Z \twoheadrightarrow (X; Y)$.
Why?

Examples

1. Fix $R(X, Y, Z)$. If $Z \to X$, then $Z \twoheadrightarrow (X; Y)$.

Why? Because $R = R_1(X, Z) \bowtie R_2(Y, Z)$ is lossless.

Examples

1. Fix $R(X, Y, Z)$. If $Z \rightarrow X$, then $Z \twoheadrightarrow (X; Y)$.
Why? Because $R = R_1(X, Z) \bowtie R_2(Y, Z)$ is lossless.

2. If $R(X, Y) = R_1(X) \times R_2(Y)$, then $R \models \emptyset \twoheadrightarrow (X; Y)$.

Examples

1. Fix $R(X, Y, Z)$. If $Z \to X$, then $Z \twoheadrightarrow (X; Y)$.
Why? Because $R = R_1(X, Z) \bowtie R_2(Y, Z)$ is lossless.

2. If $R(X, Y) = R_1(X) \times R_2(Y)$, then $R \models \emptyset \twoheadrightarrow (X; Y)$.

3. $R =$

X	Y	Z
a	x	m
a	y	m
b	x	m
b	y	m
a	x	n

Then $R \models Z \twoheadrightarrow (X; Y)$

$R_1(X, Z) =$

X	Z
a	m
b	n
a	n

$R_2(Y, Z) =$

Y	Z
x	m
y	m
x	n

Axiomatization

[Beeri et al., 1977] gave a sound and complete axiomatization for MVDs and FDs (together).

MVD1 (Reflexivity): If $Y \subseteq X$ then $X \twoheadrightarrow Y$.

MVD2 (Augmentation): If $Z \subseteq W$ and $X \twoheadrightarrow Y$ then $XW \twoheadrightarrow YZ$.

MVD3 (Transitivity): If $X \twoheadrightarrow Y$ and $Y \twoheadrightarrow Z$ then $X \twoheadrightarrow Z - Y$.

MVD4 (Pseudo-transitivity): If $X \twoheadrightarrow Y$ and $YW \twoheadrightarrow Z$ then $XW \twoheadrightarrow Z - YW$.

MVD5 (Union): If $X \twoheadrightarrow Y_1$ and $X \twoheadrightarrow Y_2$ then $X \twoheadrightarrow Y_1Y_2$.

MVD6 (Decomposition): If $X \twoheadrightarrow Y_1$ and $X \twoheadrightarrow Y_2$ then $X \twoheadrightarrow Y_1 \cap Y_2$, $X \twoheadrightarrow Y_1 - Y_2$ and $X \twoheadrightarrow Y_2 - Y_1$.

No need to read: we will see a simpler approach to MVDs

Embedded MVD

Recall that an MVD $\sigma = \boldsymbol{U} \twoheadrightarrow (\boldsymbol{V}; \boldsymbol{W})$ includes *all variables*

When σ does not include all the variables then it is called an Embedded MVD, or EMVD.

A major breakthrough:

Theorem

[Herrmann, 1995] The implication problem of EMVDs is undecidable.

Discussion

- MVDs used to define the 4th Normal Form.
- MVDs are more complex and less intuitive than FDs
- FDs equivalent to a closure system, equivalent to a lattice. No such thing for MVDs.

Conditional Independence

Definition

Fix a joint probability distribution p over variables $\boldsymbol{X}$.

$\boldsymbol{V}, \boldsymbol{W}$ are independent conditioned on $\boldsymbol{U}$ if $\forall \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$:

$$p(\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{V} = \boldsymbol{v})p(\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{W} = \boldsymbol{w}) = p(\boldsymbol{U} = \boldsymbol{u})p(\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{V} = \boldsymbol{v}, \boldsymbol{W} = \boldsymbol{w})$$

Definition

Fix a joint probability distribution p over variables $\boldsymbol{X}$.

$\boldsymbol{V}, \boldsymbol{W}$ are independent conditioned on $\boldsymbol{U}$ if $\forall \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$:

$$p(\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{V} = \boldsymbol{v})p(\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{W} = \boldsymbol{w}) = p(\boldsymbol{U} = \boldsymbol{u})p(\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{V} = \boldsymbol{v}, \boldsymbol{W} = \boldsymbol{w})$$

$$\boxed{\boldsymbol{V} \perp \boldsymbol{W} | \boldsymbol{U}} \text{ if } \boxed{p(\boldsymbol{V}, \boldsymbol{W} | \boldsymbol{U}) = p(\boldsymbol{V} | \boldsymbol{U}) \cdot p(\boldsymbol{W} | \boldsymbol{U})}$$

but be careful when $p(\boldsymbol{U} = \boldsymbol{u}) = 0$.

Definition

Fix a joint probability distribution p over variables $\boldsymbol{X}$.

$\boldsymbol{V}, \boldsymbol{W}$ are independent conditioned on $\boldsymbol{U}$ if $\forall \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$:

$$p(\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{V} = \boldsymbol{v})p(\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{W} = \boldsymbol{w}) = p(\boldsymbol{U} = \boldsymbol{u})p(\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{V} = \boldsymbol{v}, \boldsymbol{W} = \boldsymbol{w})$$

$$\boxed{\boldsymbol{V} \perp \boldsymbol{W} | \boldsymbol{U}} \text{ if } \boxed{p(\boldsymbol{V}, \boldsymbol{W} | \boldsymbol{U}) = p(\boldsymbol{V} | \boldsymbol{U}) \cdot p(\boldsymbol{W} | \boldsymbol{U})}$$

but be careful when $p(\boldsymbol{U} = \boldsymbol{u}) = 0$.

X	Y	p
0	0	1/6
0	1	1/6
1	0	1/3
1	1	1/3

$X \perp Y$?:

Definition

Fix a joint probability distribution p over variables $\boldsymbol{X}$.

$\boldsymbol{V}, \boldsymbol{W}$ are independent conditioned on $\boldsymbol{U}$ if $\forall \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$:

$$p(\boldsymbol{U}=\boldsymbol{u}, \boldsymbol{V}=\boldsymbol{v})p(\boldsymbol{U}=\boldsymbol{u}, \boldsymbol{W}=\boldsymbol{w}) = p(\boldsymbol{U}=\boldsymbol{u})p(\boldsymbol{U}=\boldsymbol{u}, \boldsymbol{V}=\boldsymbol{v}, \boldsymbol{W}=\boldsymbol{w})$$

$$\boxed{\boldsymbol{V} \perp \boldsymbol{W} | \boldsymbol{U}} \text{ if } \boxed{p(\boldsymbol{V}, \boldsymbol{W}|\boldsymbol{U}) = p(\boldsymbol{V}|\boldsymbol{U}) \cdot p(\boldsymbol{W}|\boldsymbol{U})}$$

but be careful when $p(\boldsymbol{U}=\boldsymbol{u}) = 0$.

X	Y	p
0	0	1/6
0	1	1/6
1	0	1/3
1	1	1/3

$X \perp Y$?: Yes

X	p
0	1/3
1	2/3

$\times$

Y	p
0	1/2
1	1/2

Definition

Fix a joint probability distribution p over variables $\boldsymbol{X}$.

$\boldsymbol{V}, \boldsymbol{W}$ are independent conditioned on $\boldsymbol{U}$ if $\forall \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$:

$$p(\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{V} = \boldsymbol{v})p(\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{W} = \boldsymbol{w}) = p(\boldsymbol{U} = \boldsymbol{u})p(\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{V} = \boldsymbol{v}, \boldsymbol{W} = \boldsymbol{w})$$

$$\boxed{\boldsymbol{V} \perp \boldsymbol{W} | \boldsymbol{U}} \text{ if } \boxed{p(\boldsymbol{V}, \boldsymbol{W} | \boldsymbol{U}) = p(\boldsymbol{V} | \boldsymbol{U}) \cdot p(\boldsymbol{W} | \boldsymbol{U})}$$

but be careful when $p(\boldsymbol{U} = \boldsymbol{u}) = 0$.

X	Y	p
0	0	1/6
0	1	1/6
1	0	1/3
1	1	1/3

$X \perp Y$?: Yes

X	p
0	1/3
1	2/3

$\times$

Y	p
0	1/2
1	1/2

$X \perp Y$?

X	Y	p
0	0	1/2
0	1	1/3
1	0	1/6

Definition

Fix a joint probability distribution p over variables $\boldsymbol{X}$.

$\boldsymbol{V}, \boldsymbol{W}$ are independent conditioned on $\boldsymbol{U}$ if $\forall \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$:

$$p(\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{V} = \boldsymbol{v}) p(\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{W} = \boldsymbol{w}) = p(\boldsymbol{U} = \boldsymbol{u}) p(\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{V} = \boldsymbol{v}, \boldsymbol{W} = \boldsymbol{w})$$

$$\boxed{\boldsymbol{V} \perp \boldsymbol{W} | \boldsymbol{U}} \text{ if } \boxed{p(\boldsymbol{V}, \boldsymbol{W} | \boldsymbol{U}) = p(\boldsymbol{V} | \boldsymbol{U}) \cdot p(\boldsymbol{W} | \boldsymbol{U})}$$

but be careful when $p(\boldsymbol{U} = \boldsymbol{u}) = 0$.

X	Y	p
0	0	1/6
0	1	1/6
1	0	1/3
1	1	1/3

$X \perp Y$?: Yes

X	p
0	1/3
1	2/3

$\times$

Y	p
0	1/2
1	1/2

$X \perp Y$?

X	Y	p
0	0	1/2
0	1	1/3
1	0	1/6

NO

Definition

Fix a joint probability distribution p over variables $\boldsymbol{X}$.

$\boldsymbol{V}, \boldsymbol{W}$ are independent conditioned on $\boldsymbol{U}$ if $\forall \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}$:

$$p(\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{V} = \boldsymbol{v})p(\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{W} = \boldsymbol{w}) = p(\boldsymbol{U} = \boldsymbol{u})p(\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{V} = \boldsymbol{v}, \boldsymbol{W} = \boldsymbol{w})$$

$$\boxed{\boldsymbol{V} \perp \boldsymbol{W} | \boldsymbol{U}} \text{ if } \boxed{p(\boldsymbol{V}, \boldsymbol{W} | \boldsymbol{U}) = p(\boldsymbol{V} | \boldsymbol{U}) \cdot p(\boldsymbol{W} | \boldsymbol{U})}$$

but be careful when $p(\boldsymbol{U} = \boldsymbol{u}) = 0$.

X	Y	p
0	0	1/6
0	1	1/6
1	0	1/3
1	1	1/3

$X \perp Y$?: Yes

X	p
0	1/3
1	2/3

$\times$

Y	p
0	1/2
1	1/2

$X \perp Y$?

X	Y	p
0	0	1/2
0	1	1/3
1	0	1/6

NO

Observation: if $\boldsymbol{V} \perp \boldsymbol{W} | \boldsymbol{U}$ holds then $\boldsymbol{U} \twoheadrightarrow (\boldsymbol{V}; \boldsymbol{W})$.

The Conditional Independence Implication Problem

Introduced by Pearl in the early 80s.
Given a set of CIs Σ and a CI σ, does $\Sigma \models \sigma$ hold?

[Geiger and Pearl, 1993] complete axiomatization for "saturated" CIs (meaning: each CI includes all variables).

Is the CI implication problem decidable?

Open problem for decades. There were two independent claims of proofs last year (I don't know their status).

Discussion

There is an uneasy connection between MVDs and CIs:

- MVDs correspond only to *saturated* CIs, i.e. all variables. The implication problem is the same.
- EMVDs appear to correspond to general CIs, but their implication problem is different.

Connection to Entropy

Entropic Vectors

Fix a relation instance R. [Lee, 1987] observed the following:
Let p be any probability distribution with support R, and h be its entropic vector.

For any p, $R \models \boldsymbol{U} \rightarrow \boldsymbol{V}$ iff $h(\boldsymbol{V}|\boldsymbol{U}) = 0$

Entropic Vectors

Fix a relation instance R. [Lee, 1987] observed the following:
Let p be any probability distribution with support R, and h be its entropic vector.

For any p, $R \models \boldsymbol{U} \rightarrow \boldsymbol{V}$ iff $h(\boldsymbol{V}|\boldsymbol{U}) = 0$

If p is uniform, then $R \models \boldsymbol{U} \twoheadrightarrow (\boldsymbol{V}; \boldsymbol{W})$ iff $\boldsymbol{V} \perp \boldsymbol{W} | \boldsymbol{U}$ iff $I_h(\boldsymbol{V}; \boldsymbol{W}|\boldsymbol{U}) = 0$.

Entropic Vectors

Fix a relation instance R. [Lee, 1987] observed the following:
Let p be any probability distribution with support R, and h be its entropic vector.

For any p, $R \models \boldsymbol{U} \rightarrow \boldsymbol{V}$ iff $h(\boldsymbol{V}|\boldsymbol{U}) = 0$

If p is uniform, then $R \models \boldsymbol{U} \twoheadrightarrow (\boldsymbol{V}; \boldsymbol{W})$ iff $\boldsymbol{V} \perp \boldsymbol{W} | \boldsymbol{U}$ iff $I_h(\boldsymbol{V}; \boldsymbol{W}|\boldsymbol{U}) = 0$.

X	Y	p
0	0	1/4
0	1	1/4
1	0	1/4
1	1	1/4

then

$$Z \twoheadrightarrow (X; Y)$$
$$X \perp Y | Z.$$

Entropic Vectors

Fix a relation instance R. [Lee, 1987] observed the following:
Let p be any probability distribution with support R, and h be its entropic vector.

For any p, $R \models \boldsymbol{U} \rightarrow \boldsymbol{V}$ iff $h(\boldsymbol{V}|\boldsymbol{U}) = 0$

If p is uniform, then $R \models \boldsymbol{U} \twoheadrightarrow (\boldsymbol{V}; \boldsymbol{W})$ iff $\boldsymbol{V} \perp \boldsymbol{W}|\boldsymbol{U}$ iff $I_h(\boldsymbol{V}; \boldsymbol{W}|\boldsymbol{U}) = 0$.

X	Y	p
0	0	$1/4$
0	1	$1/4$
1	0	$1/4$
1	1	$1/4$

then $Z \twoheadrightarrow (X; Y)$
$X \perp Y|Z$.

But, if probabilities are other than $1/4$, then $Z \twoheadrightarrow (X; Y)$
$\neg(X \perp Y|Z)$.

The FD/MVD implication problem can be solved with entropic inequalities!

FD/MVD Implication by Entropic Inequalities

Example: Union Axiom MVD5: $X \twoheadrightarrow Y_1, X \twoheadrightarrow Y_2 \models X \twoheadrightarrow Y_1 Y_2$

FD/MVD Implication by Entropic Inequalities

Example: Union Axiom MVD5: $X \twoheadrightarrow Y_1, X \twoheadrightarrow Y_2 \models X \twoheadrightarrow Y_1 Y_2$

Let Z be the other variables, then:

$(X \twoheadrightarrow Y_1; Y_2 Z), (X \twoheadrightarrow Y_2; Y_1 Z) \models (X \twoheadrightarrow Y_1 Y_2 | Z)$.

FD/MVD Implication by Entropic Inequalities

Example: Union Axiom MVD5: $X \twoheadrightarrow Y_1, X \twoheadrightarrow Y_2 \models X \twoheadrightarrow Y_1 Y_2$

Let Z be the other variables, then:

$(X \twoheadrightarrow Y_1; Y_2 Z), (X \twoheadrightarrow Y_2; Y_1 Z) \models (X \twoheadrightarrow Y_1 Y_2 | Z)$.

We show: $\boxed{I_h(Y_1; Y_2 Z|X) = I_h(Y_2; Y_1 Z|X) = 0 \Rightarrow I_h(Y_1 Y_2; Z|X) = 0}$

FD/MVD Implication by Entropic Inequalities

Example: Union Axiom MVD5: $X \twoheadrightarrow Y_1, X \twoheadrightarrow Y_2 \models X \twoheadrightarrow Y_1 Y_2$

Let Z be the other variables, then:

$(X \twoheadrightarrow Y_1; Y_2 Z), (X \twoheadrightarrow Y_2; Y_1 Z) \models (X \twoheadrightarrow Y_1 Y_2 | Z)$.

We show: $\boxed{I_h(Y_1; Y_2 Z | X) = I_h(Y_2; Y_1 Z | X) = 0 \Rightarrow I_h(Y_1 Y_2; Z | X) = 0}$

Suffices to show: $\boxed{I_h(Y_1; Y_2 Z | X) + I_h(Y_2; Y_1 Z | X) \geq I_h(Y_1 Y_2; Z | X)}$

Why??

FD/MVD Implication by Entropic Inequalities

Example: Union Axiom MVD5: $X \twoheadrightarrow Y_1, X \twoheadrightarrow Y_2 \models X \twoheadrightarrow Y_1 Y_2$

Let Z be the other variables, then:

$(X \twoheadrightarrow Y_1; Y_2 Z), (X \twoheadrightarrow Y_2; Y_1 Z) \models (X \twoheadrightarrow Y_1 Y_2 | Z)$.

We show: $\boxed{I_h(Y_1; Y_2 Z|X) = I_h(Y_2; Y_1 Z|X) = 0 \Rightarrow I_h(Y_1 Y_2; Z|X) = 0}$

Suffices to show: $\boxed{I_h(Y_1; Y_2 Z|X) + I_h(Y_2; Y_1 Z|X) \geq I_h(Y_1 Y_2; Z|X)}$

Why??

$$\begin{aligned}
I_h(Y_1; Y_2 Z|X) + I_h(Y_2; Y_1 Z|X) =& h(XY_1) + h(XY_2 Z) - h(XY_1 Y_2 Z) - h(X) \\
&+ h(XY_2) + h(XY_1 Z) - h(XY_1 Y_2 Z) - h(X) \\
I_h(Y_1 Y_2; Z|X) =& h(XY_1 Y_2) + h(XZ) - h(XY_1 Y_2 Z) - h(X)
\end{aligned}$$

FD/MVD Implication by Entropic Inequalities

Example: Union Axiom MVD5: $X \twoheadrightarrow Y_1, X \twoheadrightarrow Y_2 \models X \twoheadrightarrow Y_1 Y_2$

Let Z be the other variables, then:

$(X \twoheadrightarrow Y_1; Y_2 Z), (X \twoheadrightarrow Y_2; Y_1 Z) \models (X \twoheadrightarrow Y_1 Y_2 | Z)$.

We show: $\boxed{I_h(Y_1; Y_2 Z|X) = I_h(Y_2; Y_1 Z|X) = 0 \Rightarrow I_h(Y_1 Y_2; Z|X) = 0}$

Suffices to show: $\boxed{I_h(Y_1; Y_2 Z|X) + I_h(Y_2; Y_1 Z|X) \geq I_h(Y_1 Y_2; Z|X)}$

Why??

$$
\begin{aligned}
I_h(Y_1; Y_2 Z|X) + I_h(Y_2; Y_1 Z|X) =& h(XY_1) + h(XY_2 Z) - h(XY_1 Y_2 Z) - h(X) \\
&+ h(XY_2) + h(XY_1 Z) - h(XY_1 Y_2 Z) - h(X) \\
I_h(Y_1 Y_2; Z|X) =& h(XY_1 Y_2) + h(XZ) - h(XY_1 Y_2 Z) - h(X)
\end{aligned}
$$

Need to show:

$$h(XY_1) + h(XY_2 Z) + h(XY_2) + h(XY_1 Z) \geq h(XY_1 Y_2 Z) + h(X)$$

FD/MVD Implication by Entropic Inequalities

Example: Union Axiom MVD5: $X \twoheadrightarrow Y_1, X \twoheadrightarrow Y_2 \models X \twoheadrightarrow Y_1 Y_2$

Let Z be the other variables, then:

$(X \twoheadrightarrow Y_1; Y_2 Z), (X \twoheadrightarrow Y_2; Y_1 Z) \models (X \twoheadrightarrow Y_1 Y_2 | Z)$.

We show: $\boxed{I_h(Y_1; Y_2 Z|X) = I_h(Y_2; Y_1 Z|X) = 0 \Rightarrow I_h(Y_1 Y_2; Z|X) = 0}$

Suffices to show: $\boxed{I_h(Y_1; Y_2 Z|X) + I_h(Y_2; Y_1 Z|X) \geq I_h(Y_1 Y_2; Z|X)}$

Why??

$$\begin{aligned} I_h(Y_1; Y_2 Z|X) + I_h(Y_2; Y_1 Z|X) =& h(XY_1) + h(XY_2 Z) - h(XY_1 Y_2 Z) - h(X) \\ &+ h(XY_2) + h(XY_1 Z) - h(XY_1 Y_2 Z) - h(X) \\ I_h(Y_1 Y_2; Z|X) =& h(XY_1 Y_2) + h(XZ) - h(XY_1 Y_2 Z) - h(X) \end{aligned}$$

Need to show:

$$h(XY_1) + h(XY_2 Z) + h(XY_2) + h(XY_1 Z) \geq h(XY_1 Y_2 Z) + h(X)$$

Follows from $h(XY_1) + h(XY_2) \geq h(X)$ and $h(XY_2 Z) + h(XY_1 Z) \geq h(XY_1 Y_2 Z)$, which hold by modularity and non-negativity

Discussion

- Every FD/MVD implication can be derived from a Shannon inequality, where all terms are of the form $h(\boldsymbol{V}|\boldsymbol{U})$ or $I_h(\boldsymbol{V};\boldsymbol{W}|\boldsymbol{U})$ [Kenig and Suciu, 2022].
- What about general CIs? Surprisingly, there exists CIs where the conditional implication holds $I_h(\cdots) = 0 \Rightarrow I_h(\cdots) = 0$, but the corresponding inequality fails [Kaced and Romashchenko, 2013].
- Limitations of the entropic method: restricted to FD/MVDs. Next week: more general constraints, incomplete databases, probabilistic databases.

Beeri, C., Fagin, R., and Howard, J. H. (1977).
A complete axiomatization for functional and multivalued dependencies in database relations.
In *Proceedings of the 1977 ACM SIGMOD International Conference on Management of Data, Toronto, Canada, August 3-5, 1977.*, pages 47–61.

Fagin, R. (1982).
Horn clauses and database dependencies.
J. ACM, 29(4):952–985.

Geiger, D. and Pearl, J. (1993).
Logical and algorithmic properties of conditional independence and graphical models.
The Annals of Statistics, 21(4):2001–2021.

Herrmann, C. (1995).
On the undecidability of implications between embedded multivalued database dependencies.
Inf. Comput., 122(2):221–235.

Kaced, T. and Romashchenko, A. E. (2013).
Conditional information inequalities for entropic and almost entropic points.
IEEE Trans. Inf. Theory, 59(11):7149–7167.

Kenig, B. and Suciu, D. (2022).
Integrity constraints revisited: From exact to approximate implication.
Log. Methods Comput. Sci., 18(1).

Lee, T. T. (1987).
An information-theoretic analysis of relational databases - part I: data dependencies and information metric.
IEEE Trans. Software Eng., 13(10):1049–1061.