Learning and Reasoning With Incomplete Data: Foundations and Algorithms

Manfred Jaeger

Machine Intelligence Group
Aalborg University

Outline

Key References

1. D. Rubin, *Inference and Missing Data*. Biometrika 63, 1976
2. D.F. Heitjan and D. Rubin , *Ignorability and Coarse Data*. Ann. Stats. 19, 1991
3. R.D. Gill, M.J. van der Laan and J.M. Robins,
   Coarsening at Random: Characterizations, Conjectures, Counter-Examples. Proc. 1st. Seattle Symposium in Biostatistics, 1997
4. P.D. Grünwald and J.Y. Halpern, *Updating Probabilities*. JAIR 19, 2003
5. M. Jaeger, *Ignorability for Categorical Data*. Ann. Stats. 33, 2005
6. M. Jaeger, *Ignorability in Statistical and Probabilistic Inference*. JAIR 24, 2005
7. M. Jaeger, *The AI&M Procedure for Learning from Incomplete Data*. UAI 2006
8. M. Jaeger, *On Testing the Missing at Random Assumption*. ECML 2006
9. R.D. Gill and P.D. Grünwald,
   An Algorithmic and a Geometric Characterization of Coarsening at Random. Ann. Stats. 36, 2008.

Learning from Incomplete Data

heads

tails

Partially observed sequence of 10 coin tosses:

$$h, t, ?, h, ?, h, ?, h, t, ?$$

"Face-value" likelihood function for estimating the probability of *heads*:

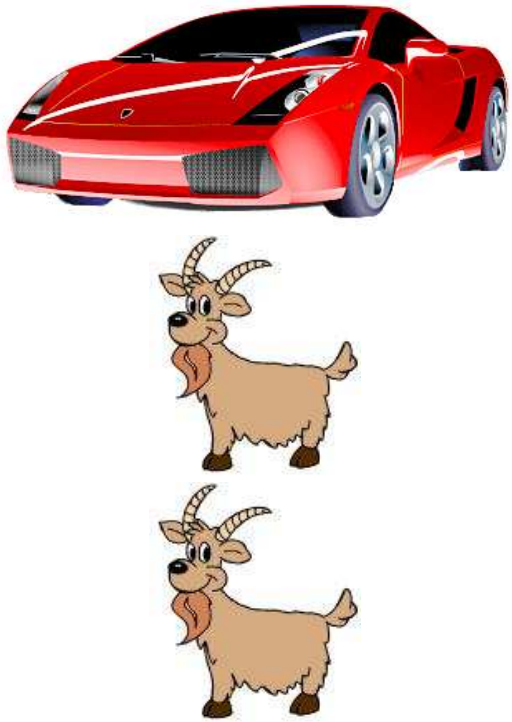
$$L(\theta) = P_\theta(data) = \prod_{i=1}^{10} P_\theta(d_i) = \theta^4 \cdot (1-\theta)^2 \cdot 1^4$$

Maximized by $\theta = 2/3$. Is this correct if "?" means: not reported because ...

- ... coin rolled off the table?
- ... one observer does not know whether "harp" is heads or tails of the Irish Euro?

Inference by Conditioning

The famous Monty Hall problem



Argument for staying with chosen door:

$$P(prize = 1 \mid prize \neq 2) = \frac{P(prize = 1)}{P(prize \in \{1, 3\})} = 1/2$$

Argument for switching to door 3:

"door 3 'inherits' the probability mass of door 2, and thus $P(prize = 3) = 2/3$"

The Common Problem

Can we identify

$$X \text{ is observed} \quad \sim \quad X \text{ has happened}$$

Coin tossing example: X: *either h or t*

Monty Hall: X: *goat behind door 2*

Outline

Missing Values and Coarse Data

Data set with missing values:

	X_1	X_2	X_3
d_1	*true*	*?*	*high*
d_2	*false*	*false*	*?*
d_3	*true*	*?*	*medium*

Other types of incompleteness:

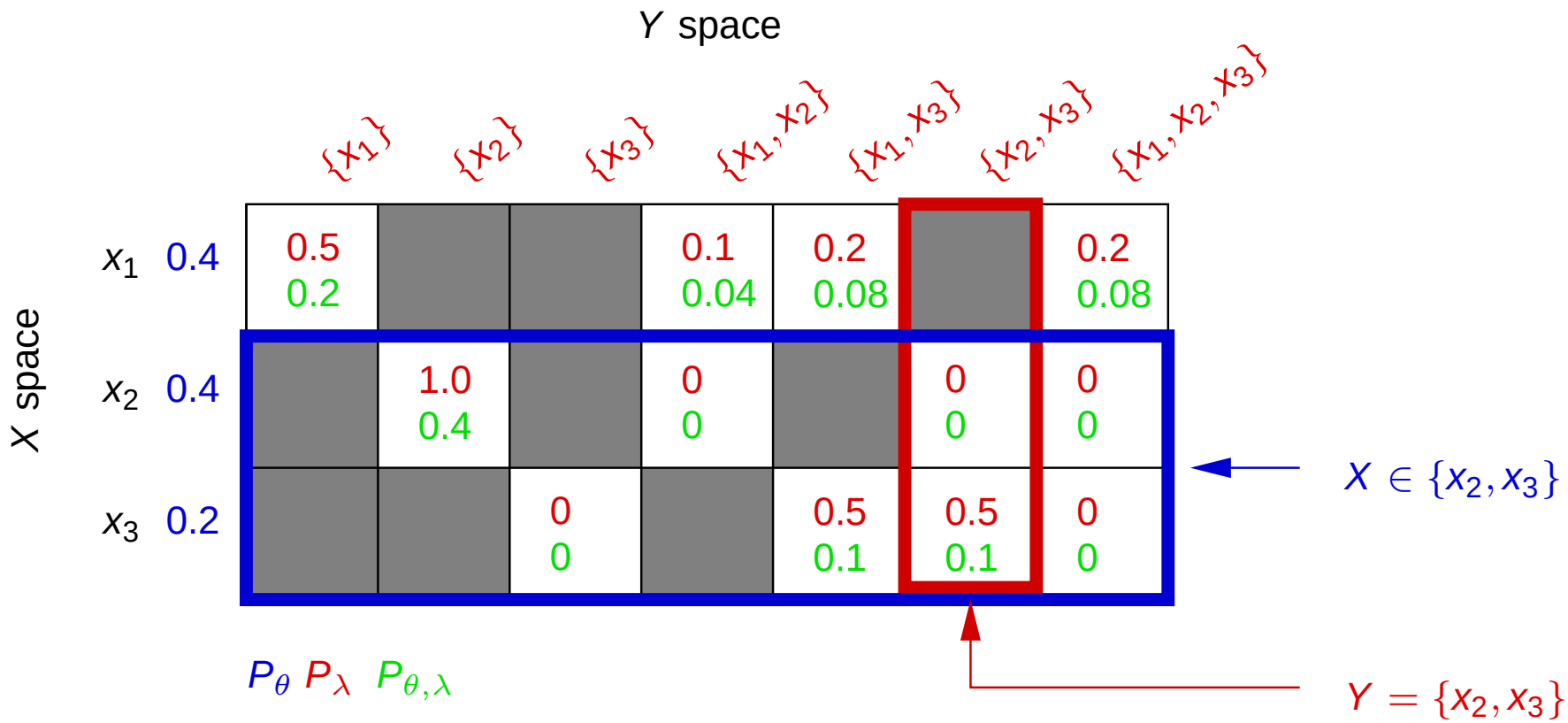
- Partly observed values: $X_3 \neq \mathit{high}$
- Constraints on multiple variables: $X_1 = \mathit{true}$ or $X_2 = \mathit{true}$

Coarse data model [2]: incomplete observations can correspond to any subset of complete observations

- More general than missing values
- Same as partial information in probability updating
  - cf. $\mathit{prize} \in \{1, 3\}$
- Simplifies theoretical analysis

Coarse Data Model

- Finite set of states (possible worlds): $W = \{x_1, \ldots, x_n\}$
- Complete data variable X with values in W, governed by distribution P_θ ($\theta \in \Theta$).
- Incomplete data variable Y with values in 2^W, governed by conditional distribution $P_\lambda(\cdot \mid X)$ ($\lambda \in \Lambda$).



Outline

Learning from Coarse Data

Data: observations of Y:

$$\boldsymbol{U} = U_1, U_2, \ldots, U_N \qquad U_i \in 2^W$$

From correct to face-value likelihood:

$$\begin{aligned}
L(\theta, \lambda \mid \boldsymbol{U}) &= \prod_i P_{\theta,\lambda}(Y = U_i) \\
&= \prod_i \sum_{x \in U_i} P_{\theta,\lambda}(Y = U_i, X = x) \\
&= \prod_i \sum_{x \in U_i} P_\theta(X = x) \, P_\lambda(Y = U_i \mid X = x) \quad \text{Ass.: constant for } x \in U_i \\
&= \prod_i P_\lambda(Y = U_i \mid X \in U_i) \sum_{x \in U_i} P_\theta(X = x) \\
&= \prod_i P_\lambda(Y = U_i \mid X \in U_i) P_\theta(U_i)
\end{aligned}$$

Profile Likelihood

$$\max_\lambda L(\theta, \lambda \mid \boldsymbol{U}) \sim \prod_i P_\theta(U_i) \quad \textit{Face-value likelihood}$$

Inference by Conditioning

Observation: value of Y:

$$U \in 2^W$$

Updating to posterior belief:

$$\begin{aligned} P_{\theta,\lambda}(X = x \mid Y = U) &= \frac{P_\theta(X = x)\, P_\lambda(Y = U \mid X = x)}{P_{\theta,\lambda}(Y = U)} \quad \text{Ass.: constant for } x \in U \\ &= \frac{P_\theta(X = x) P_\lambda(Y = U \mid X \in U)}{P_{\theta,\lambda}(Y = U)} \\ &= \frac{P_\theta(X = x) P_{\theta,\lambda}(X \in U \mid Y = U)}{P_\theta(X \in U)} \\ &= P_\theta(X = x \mid X \in U) \end{aligned}$$

Essential CAR

Data (observation) is coarsened at random (CAR) [1,2] if

$$\text{for all } U: \quad P_\lambda(Y = U \mid X = x) \quad \text{is constant for } x \in U \qquad \text{(e-CAR)}$$

The CAR assumption justifies

- learning by maximization of face-value likelihood (EM algorithm)
- belief updating by conditioning

Is that it? . . . not quite . . . what does (e-CAR) mean:

$$\text{for all } U: \quad P_\lambda(Y = U \mid X = x) \quad \text{is constant on } \{x \mid x \in U\}$$

$$\text{for all } U: \quad P_\lambda(Y = U \mid X = x) \quad \text{is constant on } \{x \mid x \in U, P_\theta(X = x) > 0\}$$

Conditioning and Weak CAR

In the justification for conditioning:

$$\frac{P_\theta(X = x)\, P_\lambda(Y = U \mid X = x)}{P_{\theta,\lambda}(Y = U)} = \frac{P_\theta(X = x) P_\lambda(Y = U \mid X \in U)}{P_{\theta,\lambda}(Y = U)}$$

Needed:

for all U : $P_\lambda(Y = U \mid X = x)$ is constant on $\{x \mid x \in U, P_\theta(X = x) > 0\}$ (w-CAR)

Profile Likelihood and Strong CAR

In the derivation of the face-value likelihood:

$$\begin{aligned}
\max_{\lambda} L(\theta, \lambda \mid \boldsymbol{U}) &= \max_{\lambda} \prod_i \sum_{x \in U_i} P_\theta(X = x) \, P_\lambda(Y = U_i \mid X = x) \\
&= \max_{\lambda} \prod_i P_\lambda(Y = U_i \mid X \in U_i) P_\theta(U_i) \\
&\approx \prod_i P_\theta(U_i)
\end{aligned}$$

- Only if domain of λ-maximization is independent of θ
  - "Paramter distinctness" [1]
- Domain of λ-maximization must not depend on $\mathit{support}(P_\theta)$
- If we assume only weak CAR, then the domain of λ-maximization *does* depend on $\mathit{support}(\theta)$
- Need

$$\text{for all } U: \quad P_\lambda(Y = U \mid X = x) \quad \text{is constant on } \{x \mid x \in U\} \qquad \text{(s-CAR)}$$

Examples

Strong CAR:

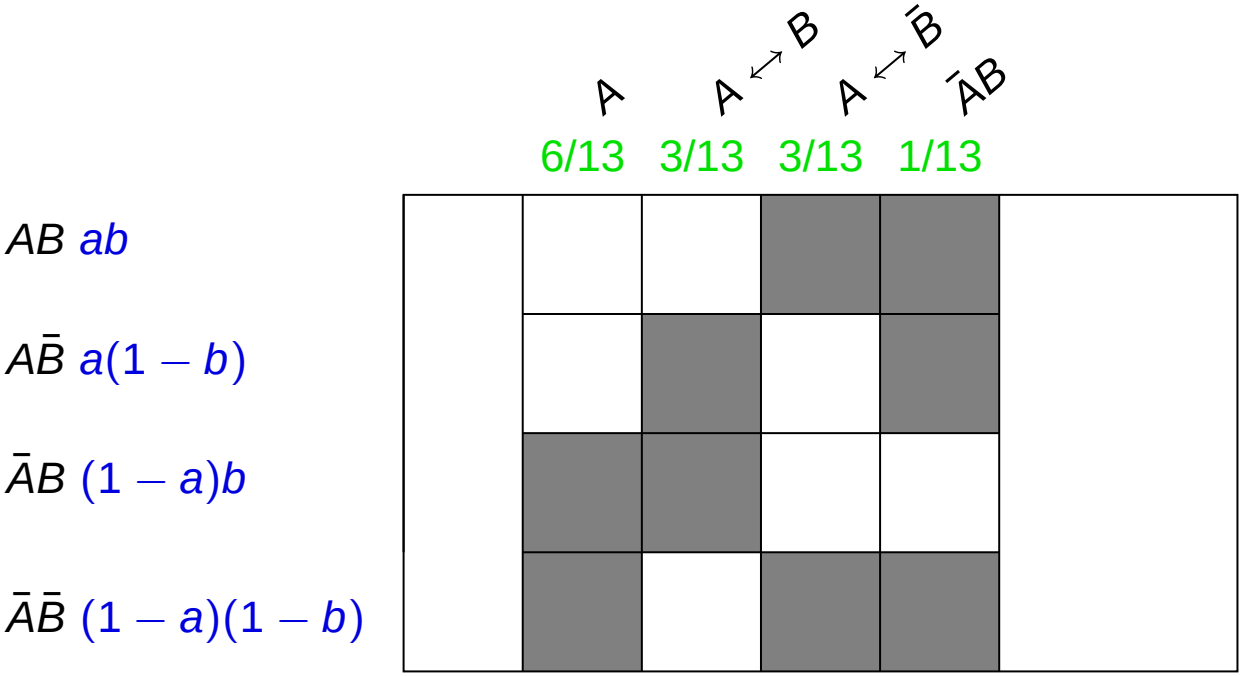
x_1	0.4	0.3			0.2	0.4		0.1
x_2	0.0		0.7		0.2		0	0.1
x_3	0.6			0.5		0.4	0	0.1

Weak CAR, not strong CAR:

x_1	0.4	0.1			0.6	0.2		0.1
x_2	0.0							
x_3	0.6			0.2		0.2	0.5	0.1

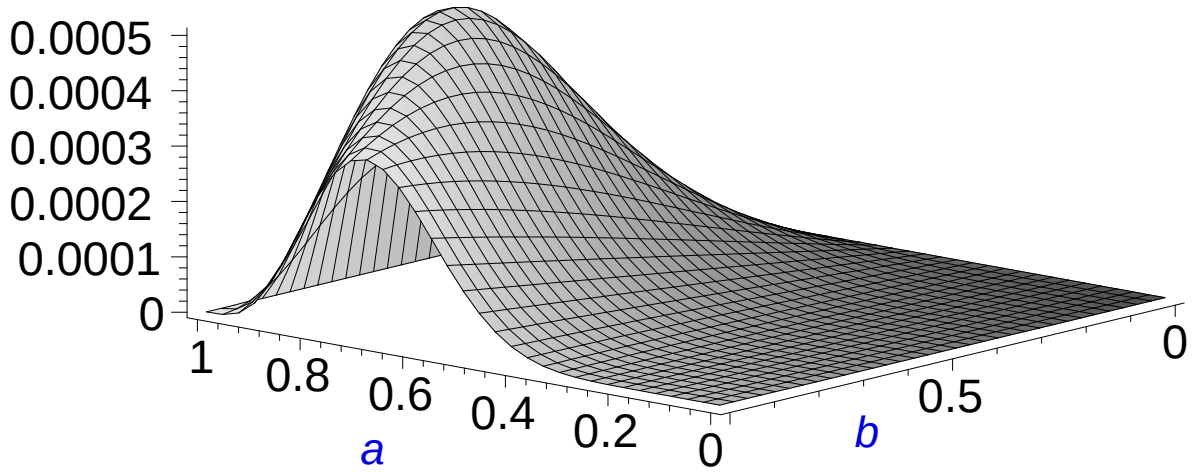
Example: Data

State space with 4 states, parametric model, and empirical probabilities from 13 observations:



Example: Face-Value Likelihood

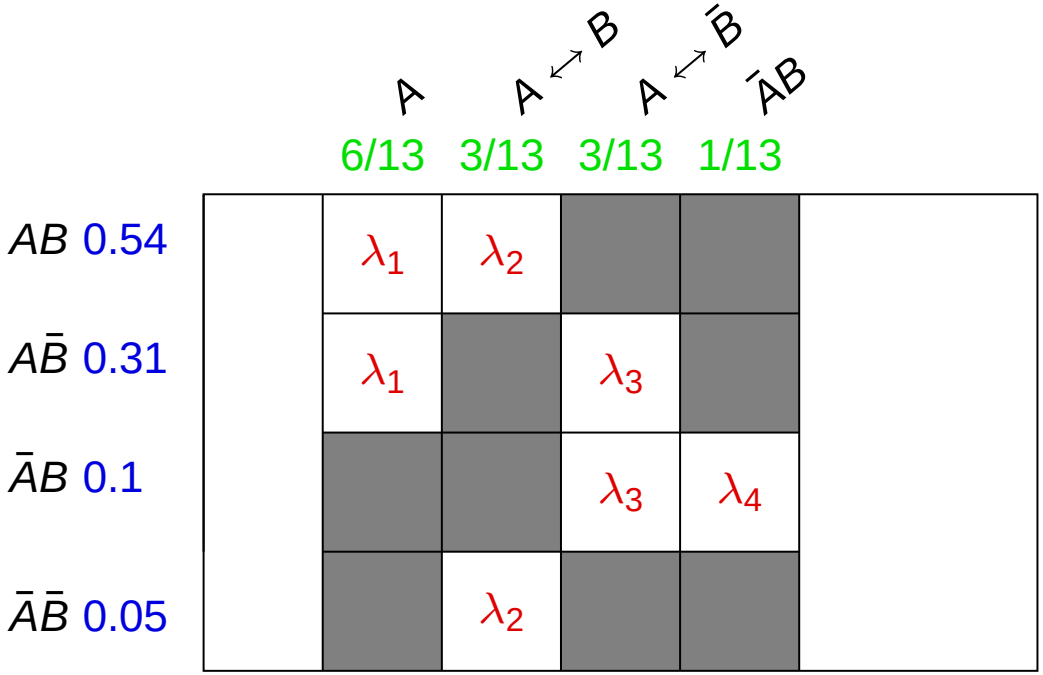
Face-value likelihood function for parameters a, b:



Maximum at $(a, b) \approx (0.845, 0.636)$

Example: Learned Distribution

Distribution learned under s-CAR assumption:

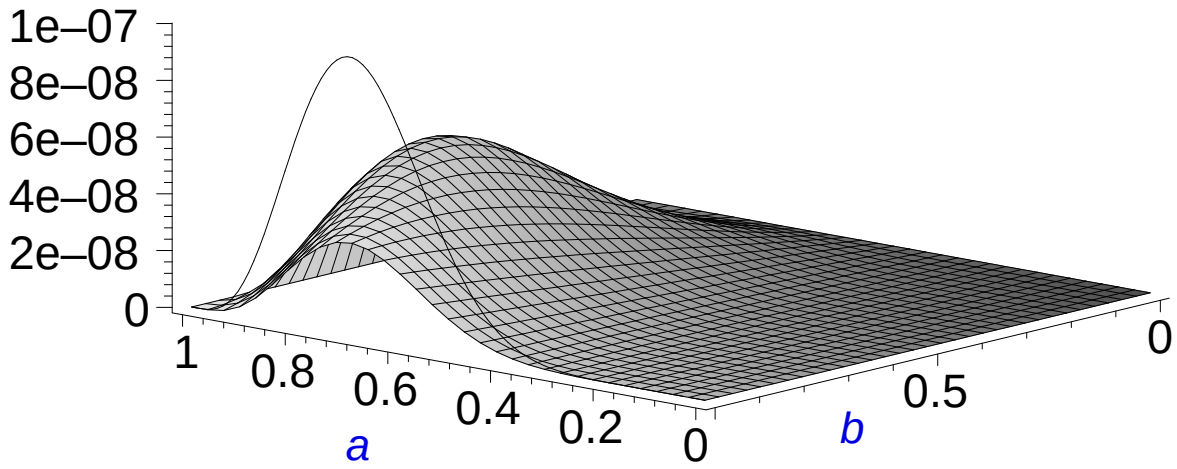


Question are there s-CAR λ parameters defining the joint distribution of X, Y with learned marginal on W, and observed empirical distribution on 2^W?

No: $\lambda_2 = 1 \Rightarrow \lambda_1 = 0 \Rightarrow P(Y = A) = 0 \neq 6/13$

Example: w-CAR Likelihood

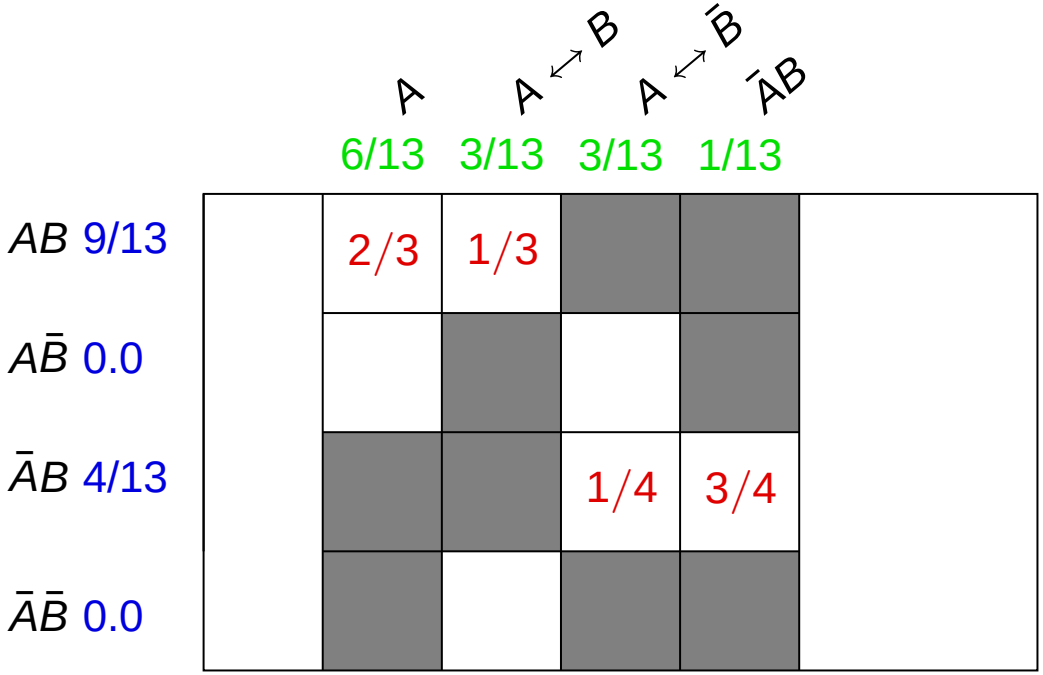
The profile likelihood under w-CAR differs from the face-value likelihood by set-of-support specific constants [5]:



Maximum at $(a, b) = (9/13, 1.0)$

Example: Learned Distribution

Distribution learned under w-CAR assumption:



Question are there w-CAR λ parameters defining the joint distribution of X, Y with learned marginal on W, and observed empirical distribution on 2^W?

Yes!

Example: Summary

The following were jointly inconsistent:

- Observed empirical distribution of Y
- Learned distribution of X under s-CAR assumption
- s-CAR assumption

Jointly consistent were:

- Observed empirical distribution of Y
- Learned distribution of X under w-CAR assumption
- w-CAR assumption

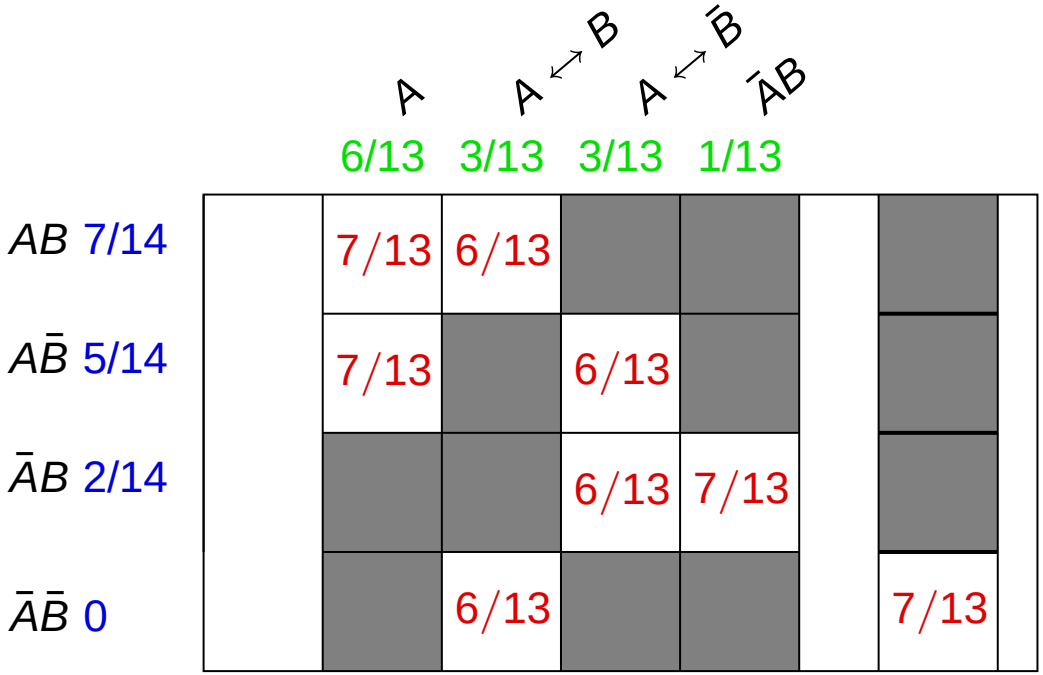
CAR is everything?

Gill, van der Laan, Robins [3]:

> *"CAR is everything"*

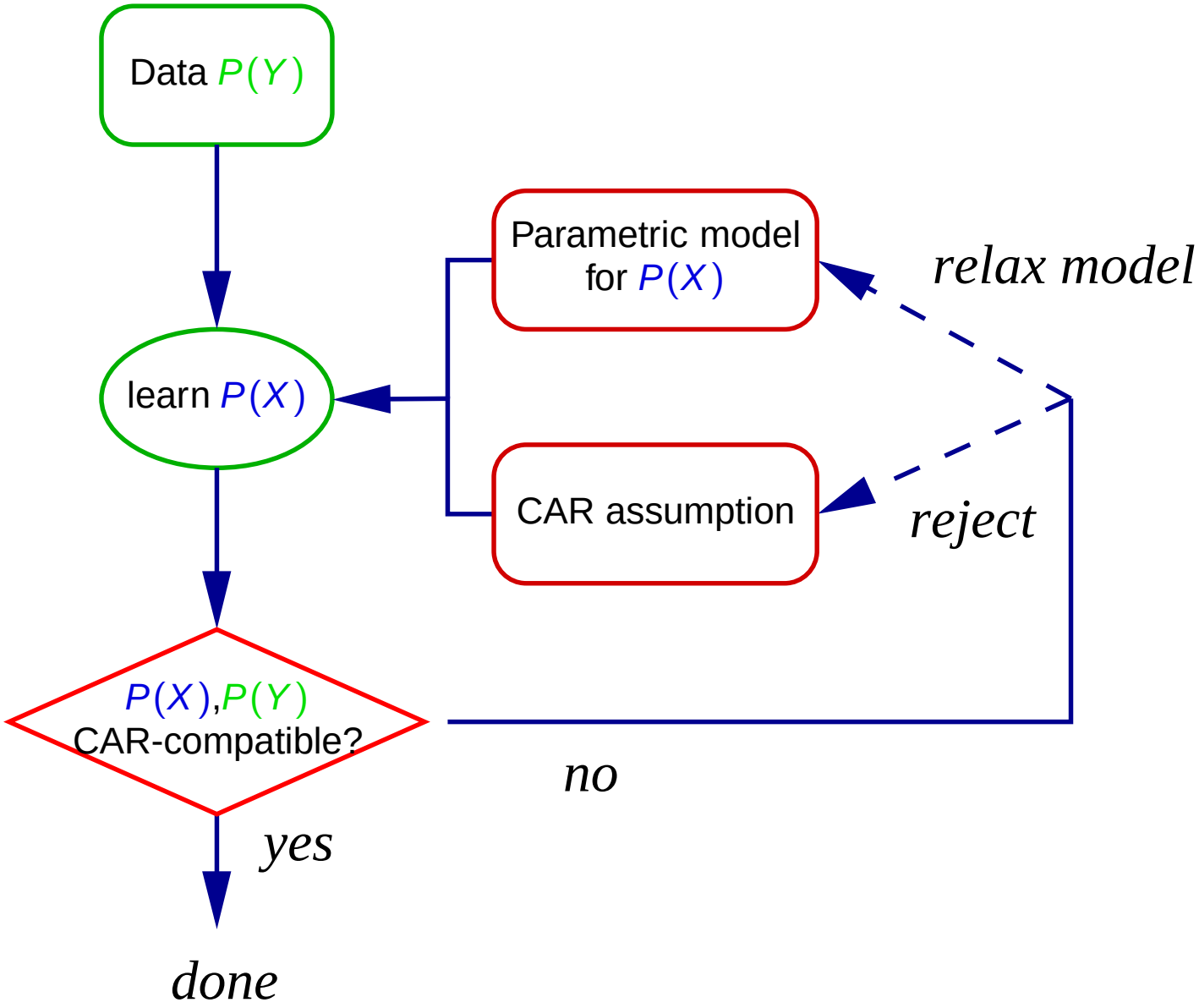
That is: for every distribution P of Y there exists a joint distribution of X, Y, s.t.

- The marginal for Y is P
- The joint is s-CAR



Outline

Testing the Assumptions



“Testing CAR”

Testing CAR relative to a parametric model:

- Set of support analysis
- Likelihood based tests

Absolute CAR “tests”:

- Compare to canonical models

Outline

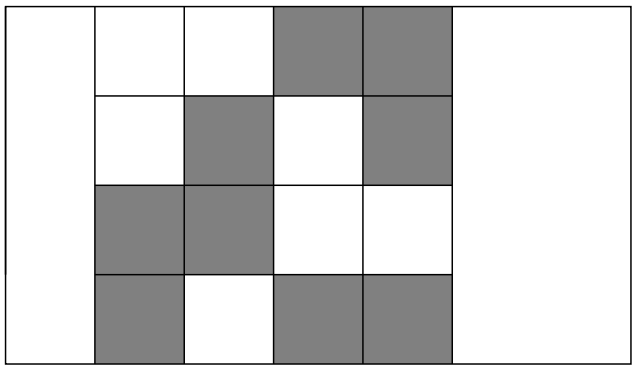
Sets of support

Assume:

- $P(X)$ has a given set of support. For simplicity: all W

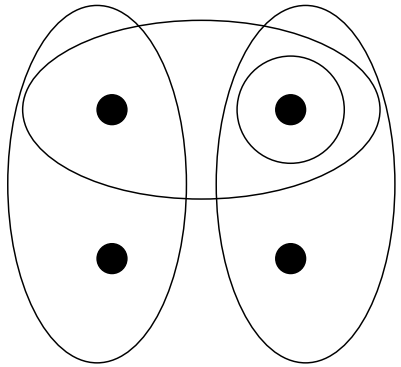
Car and sets of support

Specifications of the set of support structure (for X and Y):



$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

CARacterizing matrix [4]



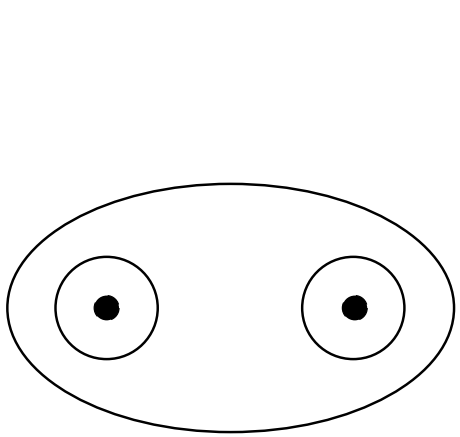
Evidence hypergraph [6]

Criteria for CAR-compatibility:

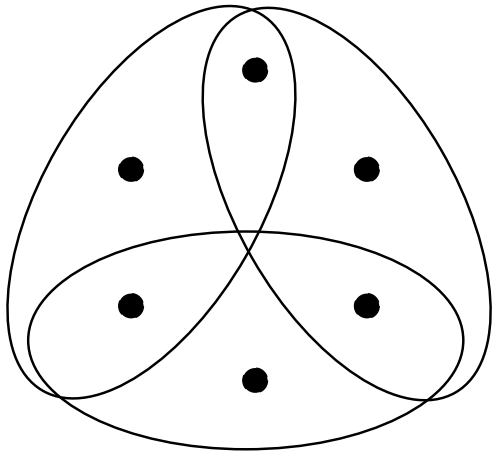
- Linear and affine relationships between rows of CARacterizing matrix
- 'Graphical' criteria on evidence hypergraph.
  - In particular: nested edges $\Rightarrow$ not CAR

Monty Hypergraphs

The evidence hypergraphs for two possible door opening scenarios (assuming fixed chosen door and 3 possible worlds: $prize = i;\ i = 1, 2, 3$):



Monty's protocol



Random opening

For Monty's protocol: hypergraph not CAR-compatible

Outline

Our Question

- Which types of data-coarsening processes or protocols generate CAR data?
- Is there a general process model that can explain all and only CAR data?

Grouped Data: Examples

Example 1

In a medical study the cigarette consumption of subjects is recorded. Initially, the data is collected in terms of *packs per day*, which is then translated into *cigarettes per day*:

1	1-20	4	0	7	21-40
2	21-40	5	1-20	8	0
3	0	6	1-20	...	...

Example 2

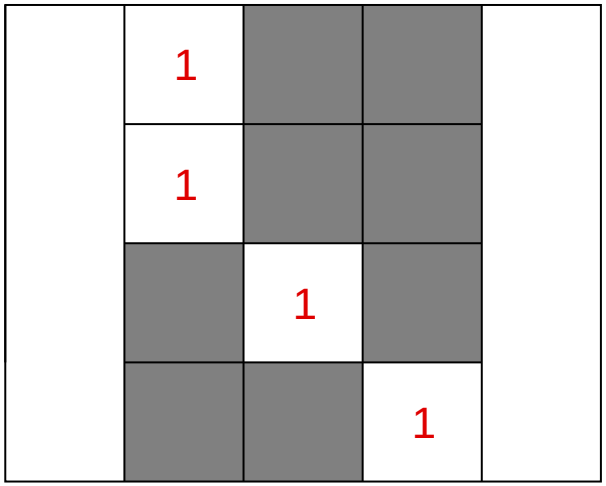
The data consists of *always observed* and *latent (never observed)* variables:

POS	?	?	?	?	?	?	...
Word	To	be	or	not	to	be	...

Grouped Data: Model

- Partition $\mathcal{W}$ of W
- $P(Y = U \mid X) = \begin{cases} 1 & \text{if } U \in \mathcal{W} \text{ and } X \in U \\ 0 & \text{otherwise} \end{cases}$

Always s-CAR:



Multiple Grouped Data: Examples

All or nothing

Each case fully observed or completely unobserved:

$$h, t, ?, h, ?, h, ?, h, t, ?$$

CAR if observed/unobserved does not depend on true value.

Cigarettes

Some protocols contain cigarette counts in # cigarettes, others in # packs:

1	12	4	0	7	21-40
2	21-40	5	7	8	0
3	0	6	1-20	...	...

CAR if cigarette count/pack count does not depend on # cigarettes.

Multiple Grouped Data: Examples

Patient Dropout

A patient's health condition is measured at the end of treatment, and at two followup examinations:

pid	*treat*	$cond_0$	$cond_1$	$cond_2$
1	y	+	+	+
2	y	+	–	–
3	n	–	–	?
4	y	–	?	?
5	n	+	+	+

CAR if dropout does not depend on treatment and initial exams.

Missing Completely at Random

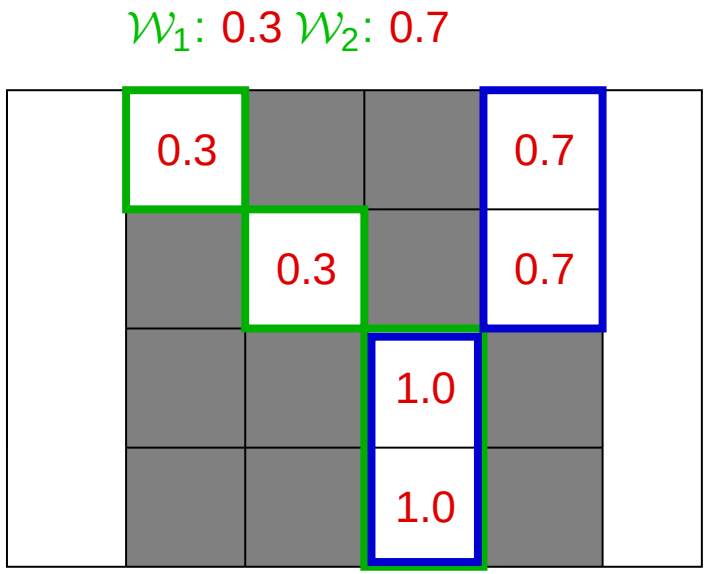
By some random process *independent of the values of* $X_1, \ldots, X_n$ it is decided which variables are observed:

id	X_1	X_2	X_3	X_4
1	t	f	?	t
2	?	f	t	f
3	f	?	?	?
4	t	t	f	f
5	?	t	f	t

Multiple Grouped Data: Model

- Partitions $\mathcal{W}_1, \ldots, \mathcal{W}_k$ of W
- Probabilities $\lambda_1, \ldots, \lambda_k$
- $P(Y = U \mid X) = \sum_{i:X \in U \in \mathcal{W}_i} \lambda_i$

Always s-CAR:



Coarsened Completely at Random

Interpretation of Multiple Grouped Data Model:

- Random choice of one of k available *sensors* or *tests*
- Report "coarse measurement" from chosen sensor/test

But beware, nonstandard sensors, not:

X	Y	X	Y
0	$\{0,1\}$	4	$\{3,4,5\}$
1	$\{0,1,2\}$	5	$\{4,5\}$
2	$\{1,2,3\}$		

CCAR

Data is *coarsened completely at random* if it is generated by a Multiple Grouped Data Model

Canonical Models 1.0

Proposals for canonical models for CAR processes:

- **Randomized Monotone Coarsening** [3]
- **CARgen** [4]
- Both equivalent to CCAR

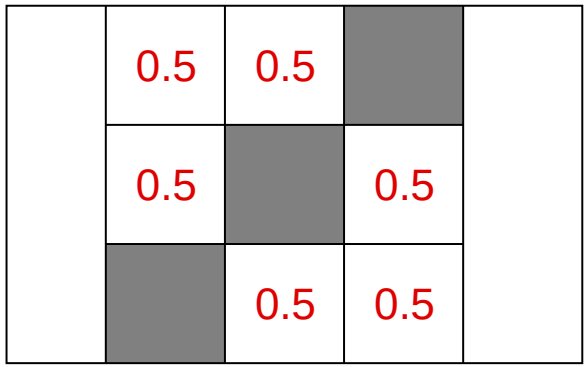
CAR models not generated by MGD process:

Weak CAR, not strong CAR:

0.1			0.6	0.2		0.1
		0.2		0.2	0.5	0.1

Beyond MGD

Strong CAR, not MGD:



A Challenge. . .

The authors cannot conceive of a more general mechanism than a randomized monotone coarsening scheme for constructing the CAR mechanisms which one would expect to meet with in practice, but is this just a lack of imagination? (Gill et al., 1997)

Uniform Noise Model

Generating U conditional on X:

$U = \{X\}$
for $i = 1, \ldots, m$
  addnoise = random boolean
  if *addnoise* = *true*
    $x =$ random uniformly selected from W
    $U = U \cup \{x\}$
return U

- Generates s-CAR data
- Can generate non-MGD
- Can not generate all s-CAR data
- Relies on uniform sampling over W

Propose and Test Model

Two solutions for a general (s-) CAR process model:

- CARgen* [4]
- Propose & Test [6]

Propose & Test

repeat until *success*
  sample $U \subseteq W$ according to Q
  if $X \in U$
    success=true
    return U

Equivalent are

- Data generated by P & T process where

$$\sum_{U:x \in U} Q(U) \text{ constant on } \{x \in W \mid P(X = x) > 0\}$$

- Data w-CAR

Robustness

MGD:
Select partition according to any probabilities $\lambda_1, \ldots, \lambda_k$

Uniform Noise:
x = random uniformly selected from W

Propose & Test:
sample $U \subseteq W$ according to Q, where $Q \ldots$

- A similar parameter condition is included in CARgen* model.

Robust procedures

A class of coarsening procedures is *robust*, if it is only defined by its qualitative protocol, but not by constraints of its probabilistic parameters.

Robustness and CAR

Equivalent are [6]:

- A CAR procedure is robust
- A CAR procedure is CCAR (i.e., MGD)

Gill & Grünwald: Uniform Multicovers

Gill & Grünwald [9] construct a general CAR procedure
"*that does not require fine-tuning of parameters*":

The more complicated parameter constraints of CARgen* or P&T are replaced by

- uniform sampling over a
- complex combinatorial space (multicovers)

Outline

Profile Likelihood

Assumptions on the coarsening mechanism $\Leftrightarrow$ Constraints on the domain of optimization for the λ in profile likelihood

$$max_{\lambda}\, L(\theta, \lambda \mid \boldsymbol{U})$$

Under S-CAR assumption:

$$\max_{\lambda} L(\theta, \lambda \mid \boldsymbol{U}) \sim \prod_{i} P_{\theta}(U_i)$$

Under *No* assumptions [7]:

$$\max_{\lambda} L(\theta, \lambda \mid \boldsymbol{U}) \sim - \min_{c \in \mathcal{C}(\boldsymbol{U})} CE(P_c, P_{\theta})$$

where

- $\mathcal{C}(\boldsymbol{U})$: space of *fractional completions* of data $\boldsymbol{U}$
- P_c: empirical distribution defined by $c \in \mathcal{C}(\boldsymbol{U})$

AI&M Algorithm

Minimize $CE(P_c, P_\theta)$ by alternating:

$c_t := arg\,min_{c \in \mathcal{C}(\boldsymbol{U})} CE(P_c, P_{\theta_t})$ (AI step)
$\theta_{t+1} := arg\,min_{\theta \in \Theta} CE(P_{c_t}, P_\theta)$ (M step)

AI&M: Example

		0.45	0.05	0.1	0.4	P_c
ab	0.1	0.05	0.05			0.1
$a(1-b)$	0.4	0.4				0.4
$(1-a)b$	0.1			0.1		0.1
$(1-a)(1-b)$	0.4				0.4	0.4

$a = 0.5$ $b = 0.2$

EM: Example

		0.45	0.05	0.1	0.4	P_c
ab	0.136	0.122	0.05			0.172
$a(1-b)$	0.363	0.327				0.327
$(1-a)b$	0.136			0.1		0.1
$(1-a)(1-b)$	0.363				0.4	0.4

$a = 0.5$ $b = 0.2727$

Outline

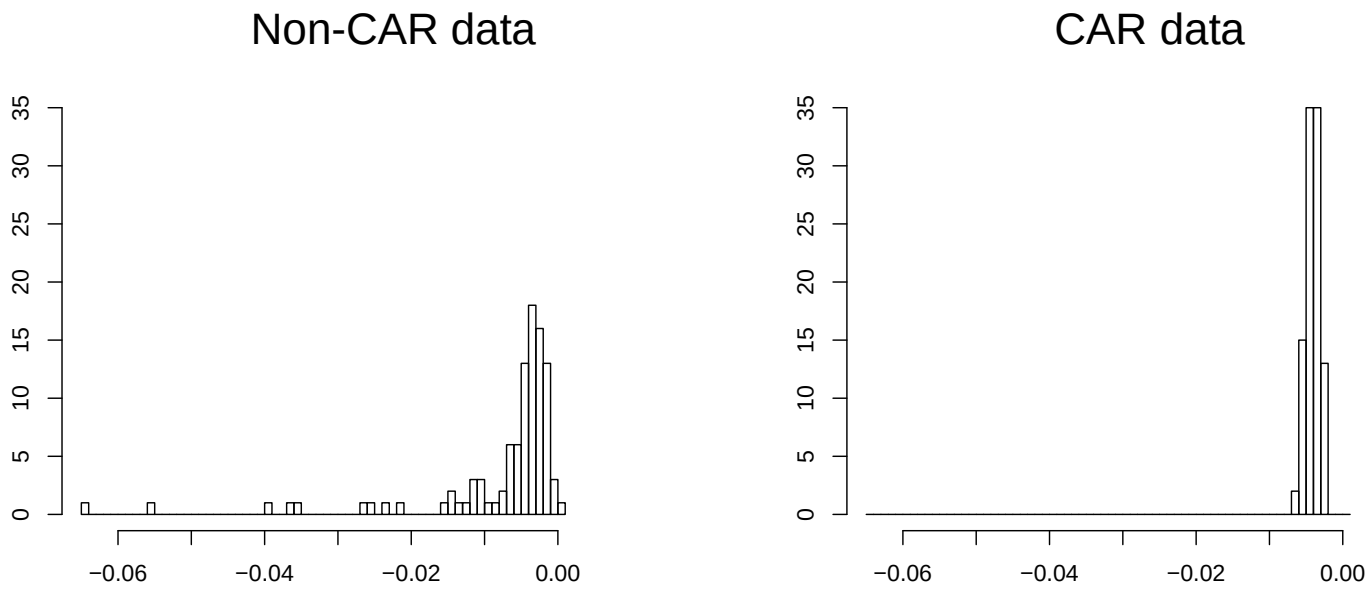
Likelihood Ratios

The AI&M learned parameters provide a better fit of the data. Reflected in the likelihood-ratio:

$$log\left(\frac{\textit{Profile-Lik-CAR-Ass(0.5,0.272)}}{\textit{Profile-Lik-No-Ass(0.5,0.2)}}\right) = -0.072$$

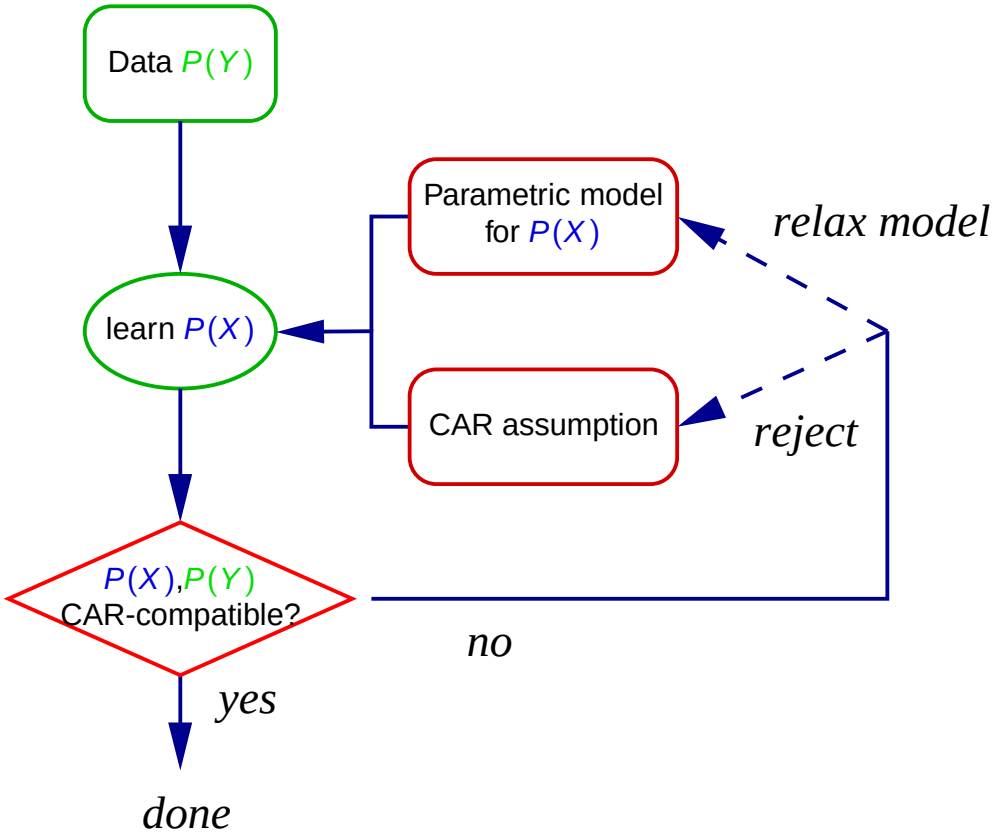
Experiment with 'Asia' network [8]

Calculated log-likelihood differences from incomplete 'Asia' data (256 states):



Testing CAR

Likelihood-ratio tests can in principle be used to test CAR *relative to a parametric model*:



Challenges:

- Computation of likelihood ratios
- Analysis of distribution of test statistic

Summary

CAR assumption instrumental for

- Learning from incomplete data with EM
- Belief updating by conditioning

“Qualitative CAR tests”:

- Support analysis
- Canonical models
  - most (all?) natural CAR models are CCAR, i.e. MGD

Learning without CAR:

- Maximize profile likelihood under NO assumptions using AI&M
- Can be the basis for quantitative statistical CAR tests.