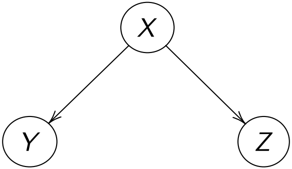
Profile Likelihood Inference in Graphical Models

Marco Cattaneo
Department of Statistics, LMU Munich

Statistische Woche 2012, Wien, Austria
18 September 2012

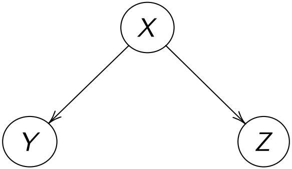
example

$X, Y, Z \in \{0, 1\}$



example

$X, Y, Z \in \{0, 1\}$

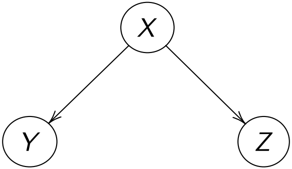


data:

X	Y	Z	#
0	0	0	15
0	0	1	25
0	1	0	7
0	1	1	5
1	0	0	6
1	0	1	35
1	1	0	3
1	1	1	4
			100

example

$X, Y, Z \in \{0, 1\}$



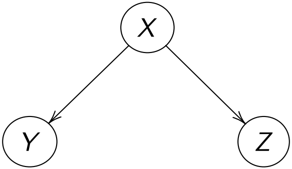
data:

X	Y	Z	#
0	0	0	15
0	0	1	25
0	1	0	7
0	1	1	5
1	0	0	6
1	0	1	35
1	1	0	3
1	1	1	4
			100

inference about $P(X = 1 \mid Y = 1,\ Z = 1)$:

example

$X, Y, Z \in \{0, 1\}$



data:

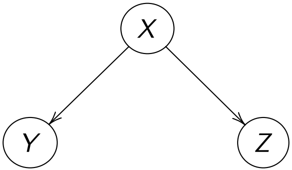
X	Y	Z	#
0	0	0	15
0	0	1	25
0	1	0	7
0	1	1	5
1	0	0	6
1	0	1	35
1	1	0	3
1	1	1	4
			100

inference about $P(X = 1 \mid Y = 1,\ Z = 1)$:

- ML estimate: 0.45

example

$X, Y, Z \in \{0, 1\}$



data:

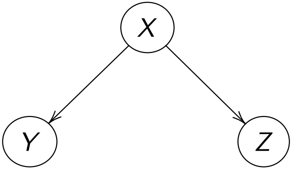
X	Y	Z	#
0	0	0	15
0	0	1	25
0	1	0	7
0	1	1	5
1	0	0	6
1	0	1	35
1	1	0	3
1	1	1	4
			100

inference about $P(X = 1 \mid Y = 1,\ Z = 1)$**:**

- ML estimate: 0.45
- Bayesian estimate
  with uniform priors: 0.46

example

$X, Y, Z \in \{0, 1\}$

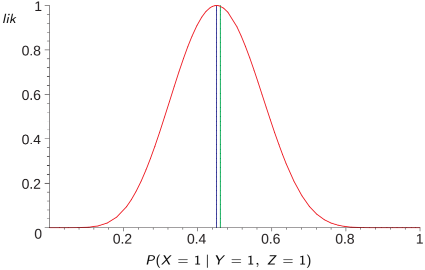


data:

X	Y	Z	#
0	0	0	15
0	0	1	25
0	1	0	7
0	1	1	5
1	0	0	6
1	0	1	35
1	1	0	3
1	1	1	4
			100

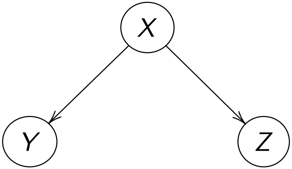
inference about $P(X = 1 \mid Y = 1,\ Z = 1)$:

- ML estimate: 0.45
- Bayesian estimate with uniform priors: 0.46
- profile likelihood function:



example $\times$ 100

$X, Y, Z \in \{0, 1\}$



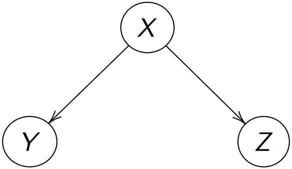
data:

X	Y	Z	#
0	0	0	1500
0	0	1	2500
0	1	0	700
0	1	1	500
1	0	0	600
1	0	1	3500
1	1	0	300
1	1	1	400
			10000

inference about $P(X = 1 \mid Y = 1,\, Z = 1)$:

example × 100

$X, Y, Z \in \{0, 1\}$



data:

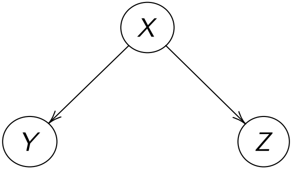
X	Y	Z	#
0	0	0	1500
0	0	1	2500
0	1	0	700
0	1	1	500
1	0	0	600
1	0	1	3500
1	1	0	300
1	1	1	400
			10000

inference about $P(X = 1 \mid Y = 1, Z = 1)$:

- ML estimate: 0.45

example $\times$ 100

$X, Y, Z \in \{0, 1\}$



data:

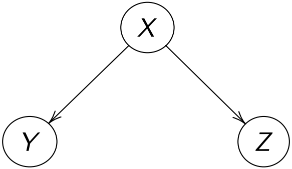
X	Y	Z	#
0	0	0	1500
0	0	1	2500
0	1	0	700
0	1	1	500
1	0	0	600
1	0	1	3500
1	1	0	300
1	1	1	400
			10000

inference about $P(X = 1 \mid Y = 1,\ Z = 1)$**:**

- ML estimate: 0.45
- Bayesian estimate
  with uniform priors: 0.46—0.01

example × 100

$X, Y, Z \in \{0, 1\}$

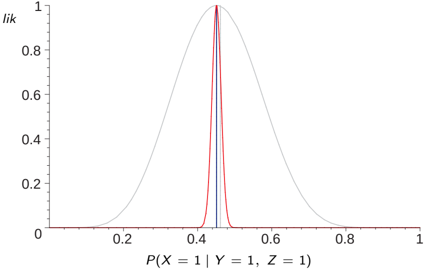


data:

X	Y	Z	#
0	0	0	1500
0	0	1	2500
0	1	0	700
0	1	1	500
1	0	0	600
1	0	1	3500
1	1	0	300
1	1	1	400
			10000

inference about $P(X = 1 \mid Y = 1, Z = 1)$:

- ML estimate: 0.45
- Bayesian estimate with uniform priors: 0.46—0.01
- profile likelihood function:



profile likelihood

- **probabilistic model**: $\{P_\theta : \theta \in \Theta\}$

profile likelihood

- **probabilistic model**: $\{P_\theta : \theta \in \Theta\}$

- **likelihood function**: $lik : \Theta \to \mathbb{R}_{\geq 0}$ with $lik(\theta) \propto P_\theta(data)$

profile likelihood

- **probabilistic model**: $\{P_\theta : \theta \in \Theta\}$

- **likelihood function**: $lik : \Theta \to \mathbb{R}_{\geq 0}$ with $lik(\theta) \propto P_\theta(data)$

- **quantity of interest:** $g(\theta)$ with $g : \Theta \to \mathbb{R}$

profile likelihood

- **probabilistic model**: $\{P_\theta : \theta \in \Theta\}$
- **likelihood function**: $lik : \Theta \to \mathbb{R}_{\geq 0}$ with $lik(\theta) \propto P_\theta(data)$
- **quantity of interest:** $g(\theta)$ with $g : \Theta \to \mathbb{R}$
- in the example: $g(\theta) = P_\theta(X = 1 \,|\, Y = 1,\, Z = 1)$

profile likelihood

- **probabilistic model**: $\{P_\theta : \theta \in \Theta\}$
- **likelihood function**: $lik : \Theta \to \mathbb{R}_{\geq 0}$ with $lik(\theta) \propto P_\theta(data)$
- **quantity of interest:** $g(\theta)$ with $g : \Theta \to \mathbb{R}$
- in the example: $g(\theta) = P_\theta(X = 1 \mid Y = 1,\ Z = 1)$
- **profile likelihood function**: $lik_g : \mathbb{R} \to \mathbb{R}_{\geq 0}$ with

$$lik_g(x) = \sup_{\theta \in \Theta \,:\, g(\theta) = x} lik(\theta)$$

basic idea

- let $f : g(\Theta) \to \mathbb{R}_{>0}$ be a strictly increasing function, and define $g' = f \circ g : \Theta \to \mathbb{R}_{>0}$

basic idea

- let $f : g(\Theta) \to \mathbb{R}_{>0}$ be a strictly increasing function, and define $g' = f \circ g : \Theta \to \mathbb{R}_{>0}$

- in the example: $f(x) = \frac{x}{1-x}$, so that $g'(\theta) = \frac{P(X{=}1, Y{=}1, Z{=}1)}{P(X{=}0, Y{=}1, Z{=}1)}$

basic idea

- let $f : g(\Theta) \to \mathbb{R}_{>0}$ be a strictly increasing function, and define $g' = f \circ g : \Theta \to \mathbb{R}_{>0}$

- in the example: $f(x) = \frac{x}{1-x}$, so that $g'(\theta) = \frac{P(X=1, Y=1, Z=1)}{P(X=0, Y=1, Z=1)}$

- for some $\alpha \in \mathbb{R}$, if θ_α **maximizes the modified likelihood function** $lik' : \Theta \to \mathbb{R}_{\geq 0}$ with

$$lik'(\theta) = lik(\theta)\, g'(\theta)^\alpha$$

basic idea

- let $f : g(\Theta) \to \mathbb{R}_{>0}$ be a strictly increasing function, and define $g' = f \circ g : \Theta \to \mathbb{R}_{>0}$

- in the example: $f(x) = \frac{x}{1-x}$, so that $g'(\theta) = \frac{P(X=1,\, Y=1,\, Z=1)}{P(X=0,\, Y=1,\, Z=1)}$

- for some $\alpha \in \mathbb{R}$, if θ_α **maximizes the modified likelihood function** $lik' : \Theta \to \mathbb{R}_{\geq 0}$ with
$$lik'(\theta) = lik(\theta)\, g'(\theta)^\alpha$$

- then the point $(g(\theta_\alpha), lik(\theta_\alpha))$ **lies on the graph of** lik_g, since
$$lik(\theta_\alpha) = \max_{\theta \in \Theta \,:\, g(\theta) = g(\theta_\alpha)} lik(\theta) = lik_g\,(g(\theta_\alpha))$$

parametric representation

- in particular (if well-defined), $\theta_0 = \hat{\theta}_{ML}$, and $\alpha \mapsto g(\theta_\alpha)$ is strictly increasing

parametric representation

- in particular (if well-defined), $\theta_0 = \hat{\theta}_{ML}$, and $\alpha \mapsto g(\theta_\alpha)$ is strictly increasing

- under regularity conditions, for some interval $\mathcal{I} \subseteq \mathbb{R}$,

$$\{(g(\theta_\alpha), lik(\theta_\alpha)) : \alpha \in \mathcal{I}\}$$

  is a **parametric representation of the graph of** lik_g

parametric representation

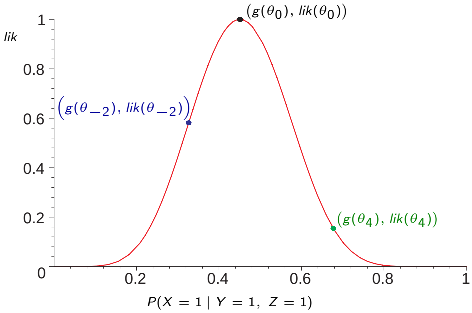
- in particular (if well-defined), $\theta_0 = \hat{\theta}_{ML}$, and $\alpha \mapsto g(\theta_\alpha)$ is strictly increasing

- under regularity conditions, for some interval $\mathcal{I} \subseteq \mathbb{R}$,

$$\{(g(\theta_\alpha), lik(\theta_\alpha)) : \alpha \in \mathcal{I}\}$$

  is a **parametric representation of the graph of** lik_g

- in the example: $\mathcal{I} = [-7, 12]$



simplest case

- in a Bayesian network with categorical variables and known graph, if the dataset is (almost) complete, then the **likelihood function factorizes**:

$$lik(\theta) = \prod_{i=1}^{m} \prod_{j=1}^{k_i} \theta_{i,j}^{n_{i,j}}, \text{ where } \sum_{j=1}^{k_i} \theta_{i,j} = 1 \text{ for all } i$$

simplest case

- in a Bayesian network with categorical variables and known graph, if the dataset is (almost) complete, then the **likelihood function factorizes**:

$$lik(\theta) = \prod_{i=1}^{m} \prod_{j=1}^{k_i} \theta_{i,j}^{n_{i,j}}, \quad \text{where} \quad \sum_{j=1}^{k_i} \theta_{i,j} = 1 \quad \text{for all} \quad i$$

- if (the f-transform of) the **quantity of interest factorizes** as well:

$$g'(\theta) = \prod_{i=1}^{m} \prod_{j=1}^{k_i} \theta_{i,j}^{q_{i,j}} \quad \text{with} \quad q_{i,j} \in \mathbb{R}$$

simplest case

- in a Bayesian network with categorical variables and known graph, if the dataset is (almost) complete, then the **likelihood function factorizes**:

$$lik(\theta) = \prod_{i=1}^{m} \prod_{j=1}^{k_i} \theta_{i,j}^{n_{i,j}}, \text{ where } \sum_{j=1}^{k_i} \theta_{i,j} = 1 \text{ for all } i$$

- if (the f-transform of) the **quantity of interest factorizes** as well:

$$g'(\theta) = \prod_{i=1}^{m} \prod_{j=1}^{k_i} \theta_{i,j}^{q_{i,j}} \text{ with } q_{i,j} \in \mathbb{R}$$

- in the example: $q_{i,j} \in \{-1, 0, 1\}$

simplest case

- then the modified likelihood function

$$lik'(\theta) = lik(\theta)\, g'(\theta)^{\alpha} = \prod_{i=1}^{m} \prod_{j=1}^{k_i} \theta_{i,j}^{n_{i,j} + \alpha\, q_{i,j}}$$

can be seen as a **likelihood function with modified data**, and is maximized by the corresponding "relative frequencies"

$$(\theta_{\alpha})_{i,j} = \frac{n_{i,j} + \alpha\, q_{i,j}}{\sum_{j'=1}^{k_i} (n_{i,j'} + \alpha\, q_{i,j'})}$$

simplest case

- then the modified likelihood function

$$lik'(\theta) = lik(\theta)\, g'(\theta)^{\alpha} = \prod_{i=1}^{m} \prod_{j=1}^{k_i} \theta_{i,j}^{n_{i,j} + \alpha\, q_{i,j}}$$

can be seen as a **likelihood function with modified data**, and is maximized by the corresponding “relative frequencies”

$$(\theta_{\alpha})_{i,j} = \frac{n_{i,j} + \alpha\, q_{i,j}}{\sum_{j'=1}^{k_i} (n_{i,j'} + \alpha\, q_{i,j'})}$$

- **parametric representation of the graph of** lik_g**:**

$$\left\{ \left(g(\theta_{\alpha}),\, lik(\theta_{\alpha}) \right) : \alpha \in \mathcal{I} \right\},$$

where $\mathcal{I} = \{\alpha \in \mathbb{R} : n_{i,j} + \alpha\, q_{i,j} \geq 0 \ \text{ for all } \ i, j\}$

classification

- **application:** Bayesian network classifier in which a class is returned only when the probabilities can be estimated with sufficient certainty

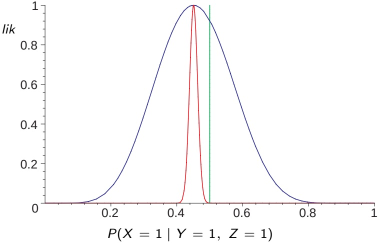
classification

- **application:** Bayesian network classifier in which a class is returned only when the probabilities can be estimated with sufficient certainty

- in the example: 0.92 and 0.00 are the degrees of uncertainty $lik_g(0.5)$ of

  $$P(X = 1 \mid Y = 1,\ Z = 1) < 0.5$$

  in the cases with 100 and 10000 data, respectively



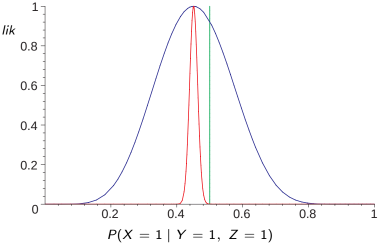
classification

- **application:** Bayesian network classifier in which a class is returned only when the probabilities can be estimated with sufficient certainty

- in the example: 0.92 and 0.00 are the degrees of uncertainty $lik_g(0.5)$ of

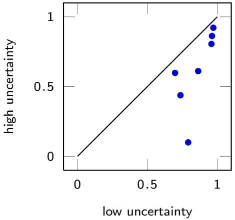
$$P(X = 1 \mid Y = 1,\ Z = 1) < 0.5$$

in the cases with 100 and 10000 data, respectively



- experimental results show that the classifier is effective in discriminating "easy" and "hard" instances

accuracy of the classification:



references

- Cattaneo (2010). **Likelihood-based inference for probabilistic graphical models: Some preliminary results**. In: *PGM 2010, Proceedings of the Fifth European Workshop on Probabilistic Graphical Models*, HIIT Publications, pp. 57–64.
- Antonucci, Cattaneo, and Corani (2011). **Likelihood-based naive credal classifier**. In: *ISIPTA '11, Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications*, SIPTA, pp. 21–30.
- Antonucci, Cattaneo, and Corani (2012). **Likelihood-based robust classification with Bayesian networks**. In: *Advances in Computational Intelligence*, Part 3, Springer, pp. 491–500.