Profile likelihood inference

Marco Cattaneo
Department of Statistics, LMU Munich

5 November 2012

example: fundamental problem of practical statistics

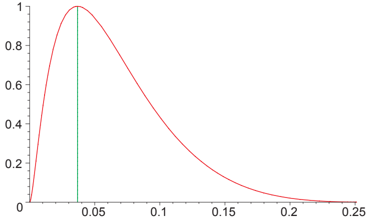
- Pearson (1920): An "event" has occurred p times out of $p+q=n$ trials, where we have no a priori knowledge of the frequency of the event in the total population of occurrences. What is the probability of its occurring r times in a further $r+s=m$ trials?

- **maximum likelihood estimate**:

$$\binom{m}{r}\left(\frac{p}{n}\right)^r\left(\frac{q}{n}\right)^s \approx 0.037$$

when $p=15,\ \ q=35,\ \ n=50,\ \ r=6,\ \ s=4,\ \ m=10$

- **uncertainty of estimate**:
  - Wald confidence interval
  - objective Bayes
  - bootstrap
  - profile likelihood



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: $\theta \in \Theta = [0, 1]$ probability of the event,

$$lik(\theta) \propto \theta^{15}\,(1-\theta)^{35}, \text{ and } g(\theta) = \binom{10}{6}\,\theta^6\,(1-\theta)^4$$

- likelihood-based confidence region for the quantity of interest:

$$\{g(\theta) : \theta \in \Theta,\ lik(\theta) > \beta\} = \{x \in \mathbb{R} :\ lik_g(x) > \beta\}$$

- **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 $g' = f \circ g : \Theta \to \mathbb{R}_{>0}$
- in the example: $f(x) = x$, so that $g' = g$
- for some $\alpha \in \mathbb{R}$, if θ_α **maximizes the modified likelihood function** $lik' : \Theta \to \mathbb{R}_{\geq 0}$ with
$$lik'(\theta) = lik(\theta)\, g'(\theta)^\alpha$$
- in the example: $lik'(\theta) \propto \theta^{15+6\,\alpha}\,(1-\theta)^{35+4\,\alpha}$ is maximized by $\theta_\alpha = \frac{15+6\,\alpha}{50+10\,\alpha}$ when $\alpha \in [-2.5, +\infty[$
- then the point $(g(\theta_\alpha), lik(\theta_\alpha))$ **lies on the graph of** lik_g, since
$$lik(\theta_\alpha) = g'(\theta_\alpha)^{-\alpha} \max_{\theta\in\Theta} lik(\theta)\, g'(\theta)^\alpha = \max_{\theta\in\Theta\,:\, g(\theta)=g(\theta_\alpha)} lik(\theta) = lik_g\,(g(\theta_\alpha))$$

parametric representation

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

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

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

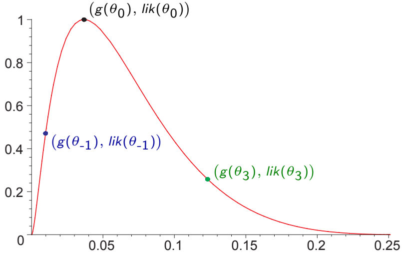
  is a **parametric representation of the graph of** lik_g

- in the example:

  $g(\theta_\alpha) = \binom{10}{6} \left(\frac{15+6\,\alpha}{50+10\,\alpha} \right)^6 \left(\frac{35+4\,\alpha}{50+10\,\alpha} \right)^4,$

  $lik(\theta_\alpha) \propto \left(\frac{15+6\,\alpha}{50+10\,\alpha} \right)^{15} \left(\frac{35+4\,\alpha}{50+10\,\alpha} \right)^{35},$

  and $\mathcal{I} = [-2.5, +\infty]$



probabilistic graphical models: simplest case

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

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

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

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

- in the example: $m = 1$, $k_1 = 2$, $\theta_{1,1} = \theta$, $\theta_{1,2} = 1 - \theta$, $n_{1,1} = 15$, $n_{1,2} = 35$, $q_{1,1} = 6$, $q_{1,2} = 4$

probabilistic graphical models: simplest case

- then the modified likelihood function

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

  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**:**

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

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

classification

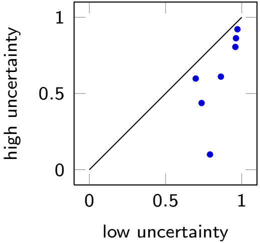
- **example of application:** Bayesian network classifier in which a class is returned only when the probabilities can be estimated with sufficient certainty

- quantity of interest:

$$P\left(C = c \mid F = f,\, (C_1, F_1) = (c_1, f_1),\, \ldots,\, (C_n, F_n) = (c_n, f_n)\right)$$

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

accuracy of the classification:



references

- 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.
- 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.
- Pearson (1920). **The fundamental problem of practical statistics**. *Biometrika* 13:1–16.