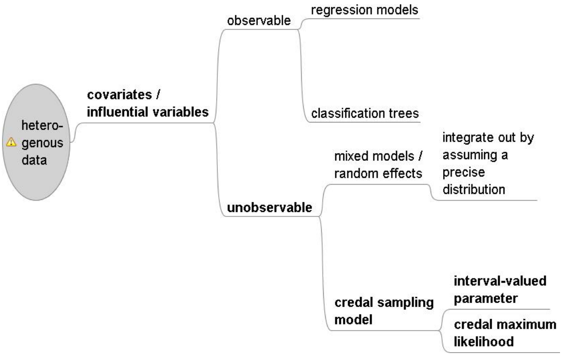
Credal maximum likelihood – an imprecise probability alternative to mixed models?

Thomas Augustin
LMU Munich

What can IP contribute to the reliable handling of unobserved data heterogeneity?



1) Traditional Maximum Likelihood

- THE frequentist estimation method
  - $\rightarrow$ consistency
  - $\rightarrow$ asymptotic normality
  - $\rightarrow$ asymptotic efficiency
  - $\rightarrow$ universally applicable
  - $\rightarrow$ gives immediately confidence regions and tests
- Observation i, $i = 1, \dots, n$
- $Y_1, \dots, Y_n := \mathbf{Y}$ outcome
- $X_1, \dots, X_n := \mathbf{X}$ covariates
- $Y_i | X_i \sim p_{vartheta, x_i}$ with density f_{ϑ, X_i}
- Estimate ϑ from observations of $Y_1, \dots, Y_n$
- After having observed $y_1, \dots, y_n$, the higher

$$\prod_{i=1}^{n} P_\vartheta(Y_i = y_i | X_i) \quad \text{or} \quad \prod_{i=1}^{n} f_{\vartheta, X_i}, \quad (*)$$

  the more plausible the conclusion that ϑ is the true parameter.
- So estimate ϑ by maximizing $(*)$ with respect to ϑ

  $\rightarrow$ *maximum likelihood estimator*

Examples

1. $Y_1, \ldots, Y_n$ normally distributed with unknown mean μ and given variance σ^2:

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{1}{2\sigma^2}(y_i - \mu)^2\right) \quad \rightarrow \quad \max_{\mu}$$

$$\iff \sum_{i=1}^{n}(y_i - \mu)^2 \quad \rightarrow \quad \min_{\mu}$$

Least square problem! (Solution $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$)

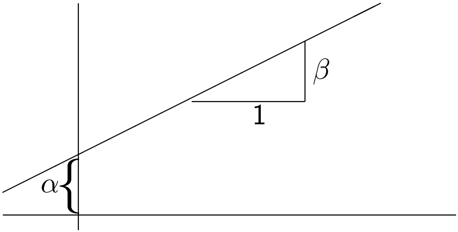
2. $Y_1, \ldots, Y_n \sim Poisson(\lambda)$

$$\prod_{i=1}^{n} \frac{\lambda^{y_i}}{y_i!} \exp(-\lambda) \quad \rightarrow \quad \max_{\lambda}$$

$$\iff \sum_{i=1}^{n}(y_i \ln \lambda - \lambda) \quad \rightarrow \quad \max_{\lambda} \Rightarrow \widehat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

3. Linear regression

$$Y_i = \alpha + \beta \cdot x_i + \varepsilon_i$$



$$y_i | x_i \sim N(x_i'\beta, \sigma^2)$$

Again maximum likelihood principle and least squares principle coincide

$$\widehat{\alpha}, \widehat{\beta} \quad \text{by} \quad \sum_{i=1}^{n} (y_i - \alpha - x_i'\beta)^2 \to \min_{\alpha,\beta}$$

2) Credal (Parametric) Sample Models

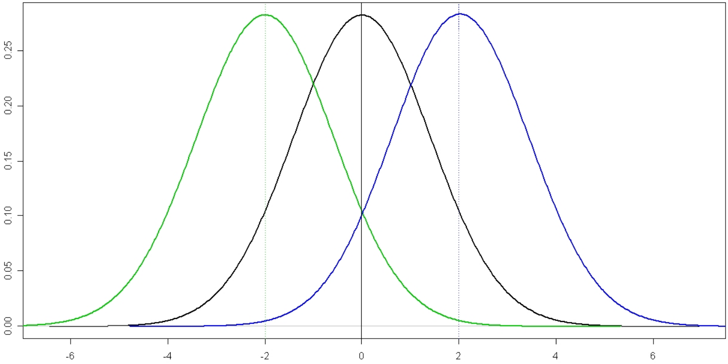
- Let $\Theta \subseteq \mathbb{R}$, parametric family of classical distributions $(p_{\vartheta,X_i})_{\vartheta\in\Theta}$.
- Credal parametric sampling model (imprecise model, not just imprecise data!).
  Parameter interval-valued
  $$[\underline{\vartheta}, \overline{\vartheta}]$$
  Credal set
  $$\mathcal{M}_{X_i} = \left\{P_{\vartheta,X_i} | \vartheta \in \left[\underline{\vartheta}, \overline{\vartheta}\right]\right\}$$
  Strongly independent observations
  $$\mathop{\mathrm{X}}_{i=1}^{n} \mathcal{M}_{X_i} = \left\{\prod_{i=1}^{n} P_{\vartheta_i,X_i} | \vartheta_i \in \left[\underline{\vartheta}, \overline{\vartheta}\right]\right\}$$

What is it good for?

Heterogeneity interpretation:
Overall parameter + individual parameter:

$$\vartheta_i = \vartheta_{overall} + \nu_i$$

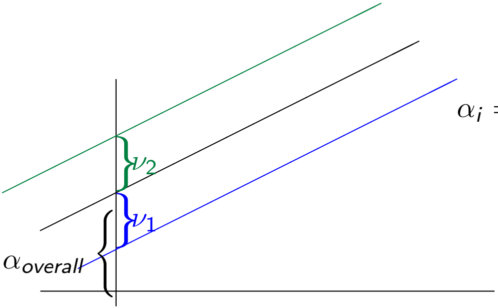
		unobserved
biometrics	overall treatment effect	hospital-, patient-specific
insurance	overall risk	individual risk attitude
dynamical econometrical model	overall chance	individual characteristics


0.00
0.05
0.10
0.15
0.20
0.25
-6
-4
-2
0
2
4
6

In linear regression analysis "set of "true" regression lines"

simple linear regression: x_i one-dimensional

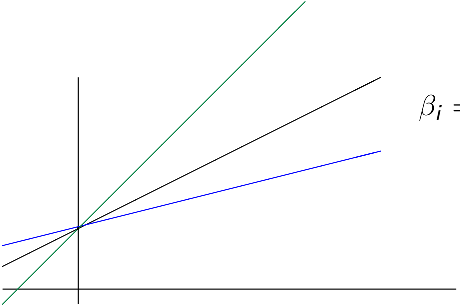
i) $y_i = \alpha_i + \beta x_i + \varepsilon$



$\alpha_i = \alpha_{overall} + \nu_i$

x_i dummy variables: Analysis of variance with random effects

ii) $y_i = \alpha + \beta_i x_i + \varepsilon$



$\beta_i = \beta_{overall} + \nu_i$

3) Traditional solution: random effects model

Assume certain distribution, described by $\tilde{f}(\cdot)$, for ν_i, typically $\nu_i \sim N(0, \sigma_\nu^2)$
Consider likelihood

$$\prod_{i=1}^{n} f_{\vartheta_{overall}}(x_i) = \prod_{i=1}^{n} \int f_{\vartheta_{overall}}(x_i|\nu_i) \cdot \tilde{f}(\nu_i) d\nu_i$$

to estimate $\vartheta_{overall}$

- point estimator, irrespectively of amount of heterogeneity
- depends, of course, strongly on $\tilde{f}(\cdot)$

4) Level δ – Credal Maximum Likelihood Estimation

Definition:

Let $\delta \geq 0$ be fixed and let, for given data $y_1, y_2, \ldots, y_n$,

$$\hat{\vartheta}_1, \ldots, \hat{\vartheta}_n, \quad \widehat{L\vartheta}, \widehat{U\vartheta},$$

be an optimal solution of

$$\prod_{i=1}^{n} f_{\vartheta_i}(y_i) \to \max_{\vartheta_1,\ldots,\vartheta_n, L\vartheta, U\vartheta}$$

subject to

$$\begin{aligned} L\vartheta \leq \vartheta_i &\leq U\vartheta, \quad i = 1, \ldots, n \\ U\vartheta - L\vartheta &\leq \delta\,, \end{aligned}$$

then

$$\left[\widehat{L\vartheta}, \widehat{U\vartheta}\right]$$

is called level-δ credal maximum likelihood estimator.

Remarks:

i) Obviously

$$\delta = 0 \Rightarrow \widehat{L\vartheta} = \widehat{U\vartheta} = \hat{\vartheta}_{ML}$$

(the traditional ML estimator)

ii) Of course, it is much more convenient to replace the objective function by the equivalent objective function

$$\sum_{i=1}^{n} \ln f_{\vartheta_i}(y_i) \to \max$$

5) Examples: Least Squares Problems

Example I: normal model: Normal distribution (ML and Least Squares coincide), parameter μ_i.

We have to consider the quadratic optimization problem

$$\sum_{i=1}^{n}(y_i - \mu_i)^2 \to \min$$

subject to

$$L\mu \leq \mu_i \leq U\mu \quad \text{and} \quad U\mu - L\mu \leq \delta,$$

which can be solved by standard software.

i) At least in the case of normal distribution with unknown location parameter

$$\delta \to \infty : \widehat{L\vartheta} = \min_{i=1,\ldots,n} y_i$$

ii) The problem can be viewed as a function of the lower interval limit T of the estimator

$$\mathcal{E}(y_i, T) = (y_i - T)^2 \cdot I\{y_i \leq T\} + \left(y_i - (T + \delta)^2\right) \cdot I\{y_i \geq T + \delta\}$$

Some numerical toy examples ($n = 4$, MAPLE)

i) $y_1 = 1;\ y_2 = 2;\ y_3 = 3;\ y_4 = 4$

δ	$\left[\widehat{L\mu}, \widehat{U\mu}\right]$
0:	2.5 ...
0.1:	[2.45; 2.55]
0.5:	[2.25; 2.75]
1:	[2; 3]

ii) Note $\left[\widehat{L\mu}, \widehat{U\mu}\right]$ is not just $\hat{\mu} \pm$ something

$y_1 = 1;\ y_2 = 2;\ y_3 = 3;\ y_4 = 14$

δ	$\left[\widehat{L\mu}, \widehat{U\mu}\right]$
0:	5
0.1:	[4.975; 5.075]
0.5:	[4.875; 5.375]
1:	[4.75; 5.75]

Example II: simple linear regression In the regression context we have to consider

$$\sum_{i=1}^{n}(y_i - \alpha_i - \beta x_i) \to \min$$

or

$$\sum_{i=1}^{n}(y_i - \alpha - \beta_i x_i) \to \min$$

subject to the restrictions

$$\alpha_i \in [\widehat{L_\alpha}, \widehat{U_\alpha}], \quad \widehat{U_\alpha} - \widehat{L_\alpha} \leq \delta$$

and

$$\beta_i \in [\widehat{L\beta}, \widehat{U\beta}], \quad \widehat{U\beta} - \widehat{L\beta} \leq \delta$$

respectively

6) Outlook

Conjecture: Objective function convex then

$$\delta_1 \leq \delta_2 \Rightarrow [\widehat{L_{\delta_1}\vartheta}, \widehat{U_{\delta_1}\vartheta}] \subseteq [\widehat{L_{\delta_2}\vartheta}, \widehat{U_{\delta_2}\vartheta}]$$

$\rightarrow$ Note special case $\delta = 0$ (tradit. ML)
Then:

Under i.i.d ($\vartheta_i \equiv \vartheta$)

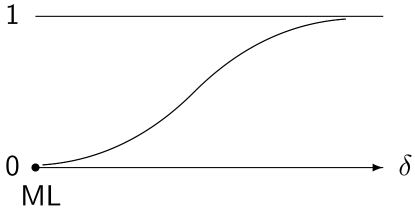
$$\lim_{n\rightarrow\infty} P_\vartheta \left(\left[\widehat{L_\delta \vartheta^{(n)}}, \widehat{U_\delta \vartheta^{(n)}} \right] \ni \vartheta \right) = 1$$

"i.i.d consistency of level δ-ML estimation"
(Proof: traditional consistency of ML; conjecture above)

On the Choice of δ

a) Look at the objective function as a function of δ:
Use that δ where a further increase does not improve the objective function substantially (cp. elbow criterion in principal component analysis)

b) fuzzy set interpretation; estimator as a fuzzy set, membership function increasing in δ



c) Penalization (like in nonparametric statistic)
look at the objective function

$$\sum_{i=1}^{n} \ln f_{\vartheta}(y_i) - \lambda \cdot \delta \to \max$$

λ for instance by cross-validation

Additional aspects

- What can we learn when $(P_\vartheta)_{\vartheta \in \Theta}$ is stochastically ordered?
- Comparison to traditional random effect models
- Method can be extended to robust objective functions $\to$ credal M-estimators