Credal Sampling Models and Credal Maximum Likelihood

Thomas Augustin

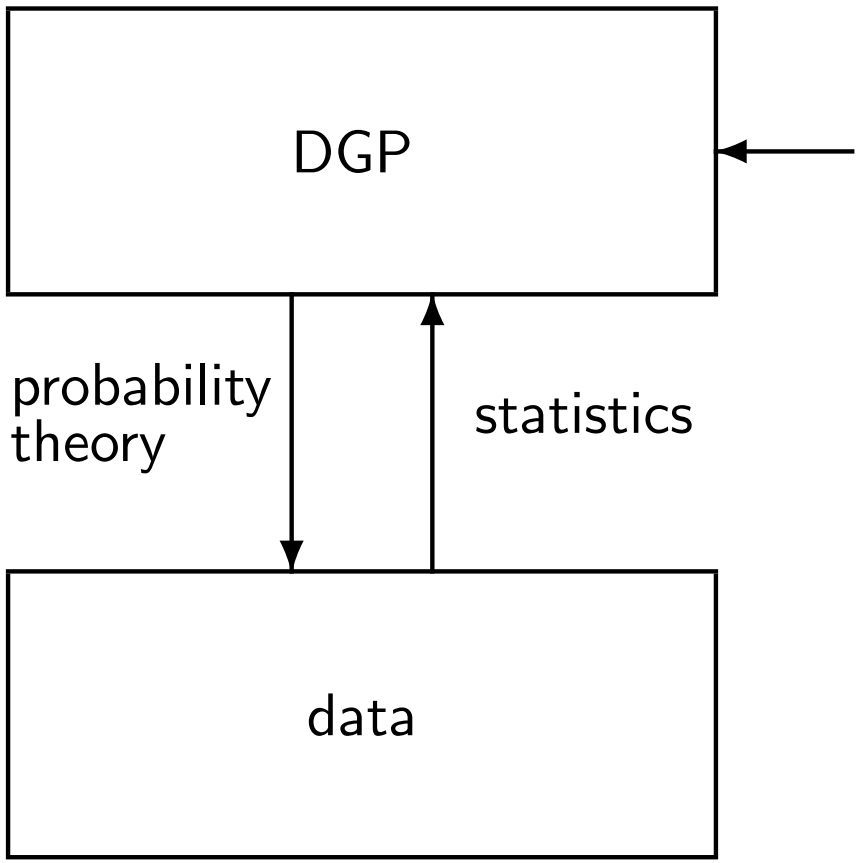
Foundations of Statistics and Their Applications Group
Department of Statistics
LMU Munich

Background and Aims

- Review some aspects of imprecise sampling models and their potential in statistics
- Work out some "missing links" as research questions
- Briefly discuss one approach $\rightarrow$ credal maximum likelihood, which may be interpreted as an imprecise probability alternative to mixed models

- "Lugano discussion": Is there such a thing like imprecise sampling (data) model. If so, can we distinguish data from "imprecise models" and from precise models?
- "Ghent and Old Library discussion": Why does IP (imprecise *probability*) have so little impact in statistics ?

- Claim: In order to utilize the full power of imprecision in statistic, imprecision has to be incorporated into the core of statistical modelling, i.e. into *sampling models*



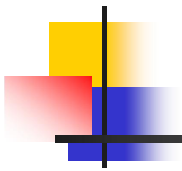
Somewhat strange situation in statistics with IP

$\rightarrow$ uncertainty about potential observations is naturally expressed by IP: $([\underline{P}(X), \overline{P(X)}]$ for gambles X,

$\rightarrow$ "inference" is typically understood as conditioning on (the outcome of) a (partially informative) gamble: $[\underline{P}(X|Y), \overline{P}(X|Y)]$ (or $[\underline{P}(X|y), \overline{P}(X|y)]$

- but great scepticism about "imprecise sampling models"

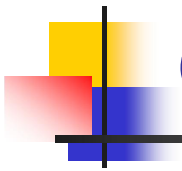
Fine (2003, Dinner Talk, ISIPTA)



Foundations Matter!

- It is not about neatly organizing what we all agree on!
- Foundational choices have real-world consequences.
- Foundations of probability are much closer to the surface of the applications of probability than are the foundations of arithmetic to accounting or the foundations of physics to physics (Suppes 98) or to engineering.

Fine (2003, Dinner Talk, ISIPTA)



Attending to the Objective/Empirical

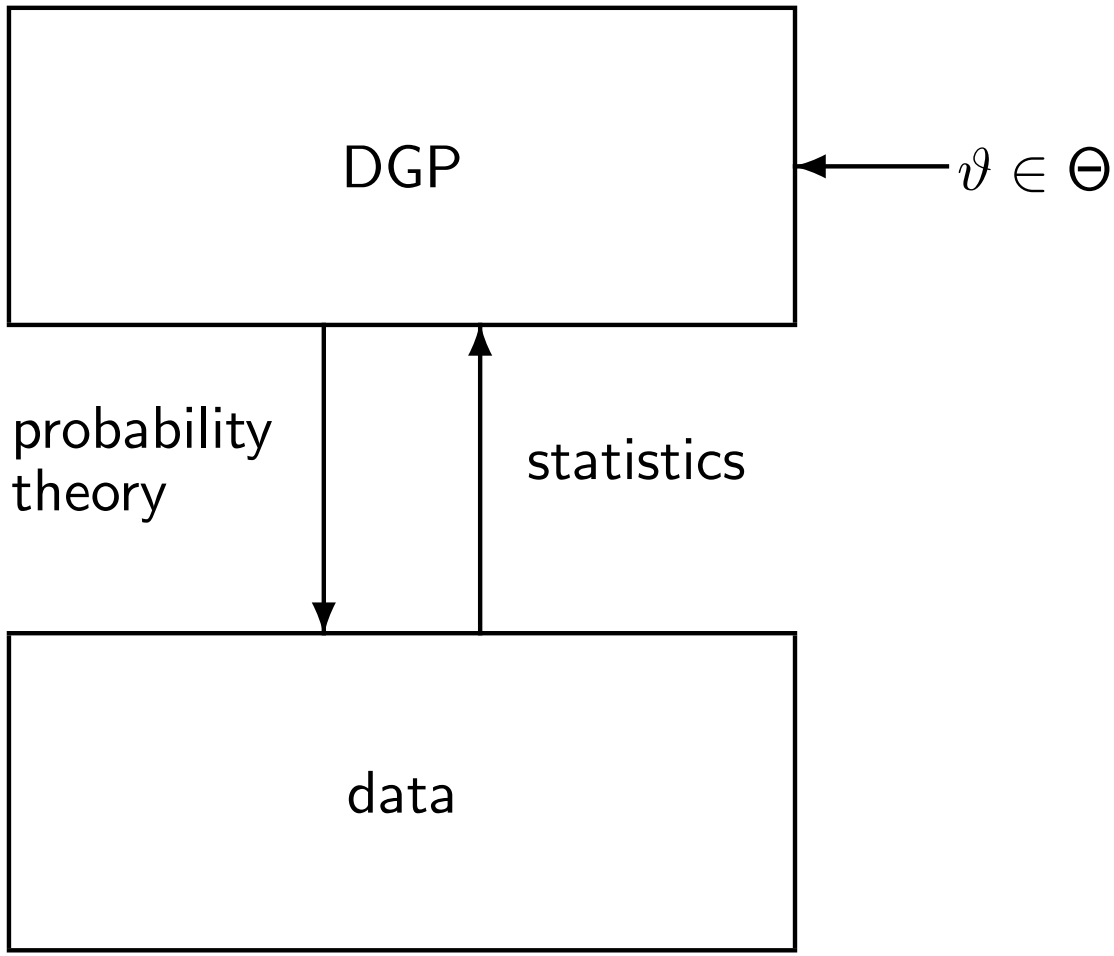
- ISIPTA and others have shown a marked preference for the subjective and personalistic view of probability.
- The subjective view profits from the mind's limited insight into the brain.
- PROBABILITY DOES EXIST
- Objective empirical probability is far too important to suffer neglect.

(Parametric) Statistical Modelling

- two kinds of entities
  - $\rightarrow$ specific observables of interest $\rightarrow$ variables $\rightarrow$ data
  - $\rightarrow$ potential "states of the world" $\rightarrow$ parameter ϑ, parameter space Θ

  "Probability model", Sampling model:

  $$P(\text{Observables} \,||\, \text{parameter})$$


DGP
$\vartheta \in \Theta$
probability
theory
statistics
data

- "classical": ϑ fixed, index of $(p_\vartheta(X))_{\vartheta\in\Theta}$
- Bayesian (reinterpreted): ϑ outcome of (latent) variable Θ
  express knowledge by specifying a prior distribution
- Statistics: learn structure from repeated observations
  - $\Rightarrow$ characterize "true" state of nature $\rightarrow$ sets of values compatible with the data ("confidence regions, hypothesis testing, credibility regions; single valued: point estimation or posterior distribution)
  - $\rightarrow$ concepts of independence become crucial (or more generally concepts how to combine individual elements for constructing the joint distribution)

Later on: Maximum Likelihood

- THE estimation method in traditional statistics
- conditional, with very good frequentist properties
  - → consistency
  - → asymptotic normality
  - → asymptotic efficiency
  - → universally applicable
  - → gives immediately confidence regions and tests

- Observation i, $i = 1, \ldots, n$
- $Y_1, \ldots, Y_n := \mathbf{Y}$ outcome
- $(X_1, \ldots, X_n := \mathbf{X})$ covariates
- $Y_i | X_i \sim p_{\vartheta, x_i}$ with density f_{ϑ, X_i}
- Estimate ϑ from observations of $Y_1, \ldots, Y_n$
- After having observed $y_1, \ldots, y_n$, the higher

$$\prod_{i=1}^{n} P_\vartheta(Y_i = y_i || X_i) \qquad (*)$$

  the more plausible is 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:

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

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

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

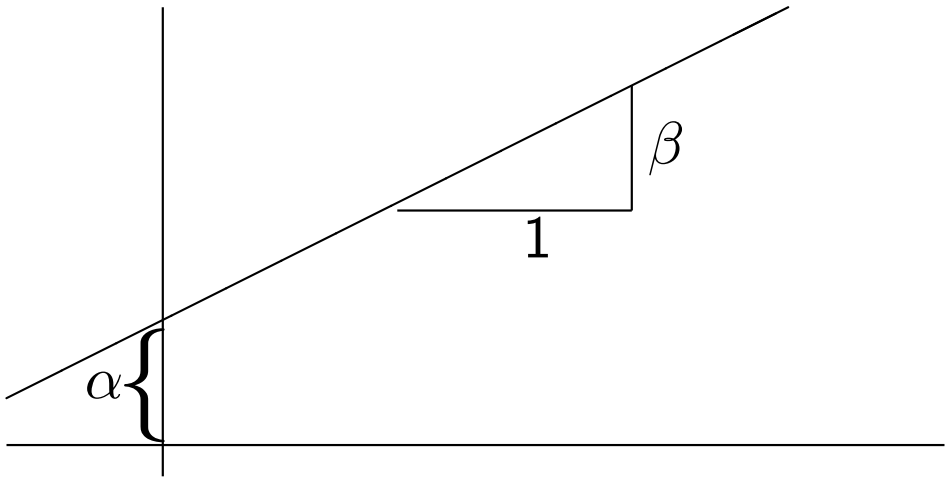
2. $Y_1, \ldots, Y_n \sim \mathit{Poisson}(\lambda)$

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

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

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

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

- traditional setting: true classical probability $p(\cdot)$ i.i.d. observations $X_1, ...X_n$, $X_i \sim p$

  $p(X_1 = x_1, X_2 = x_2, ...X_n = x_n) = \prod_{i=1}^{n} p(X_i = x_i) =: p^{\otimes n}(x_1, ..., x_n)$

- now i.i.d observations $X_1, ...X_n, X_i \sim P(\cdot) = [L(\cdot), U(\cdot)]$ with the structure (corresponding credal set) $\mathcal{P}$

  at least two settings have to be distinguished

  - strict repetition point of view
    $X_i \sim \underbrace{p}_{(!!!)} \in \mathcal{P}; i = 1, ..., n$
  - heterogeneous independence $X_i \sim \underbrace{p_i}_{(!!!)} \in \mathcal{P}; i = 1, ..., n$
  - room for creative ideas, epistemic independence?

Statistical models

- $\mathcal{P}\bigotimes n, \mathrm{rep}$ $\mathcal{P}^{\bigotimes n, het}$
  ((clearly better notation needed))
- traditional parametric model: type of distribution known up to a parameter $\vartheta \in \Theta \subseteq \mathbb{R}^q \to (p_\vartheta^{\otimes n})_{\vartheta \in \Theta}$ find true ϑ

How to transfer this to IP?

- single-valued parameter IP models $(\mathcal{P}_\vartheta^{\otimes n,\bullet})_{\vartheta\in\Theta}$ ($\bullet \in \{\text{rep, het}\}$ typically neighbourhood models:
  - central distribution $q_\vartheta(\cdot)$
  - $\mathcal{P}_\vartheta$ any distribution "close" to $q_\vartheta(\cdot)$
- interval-parameter model

$$\boldsymbol{\vartheta} = [\underline{\vartheta}, \overline{\vartheta}] \subseteq \Theta$$

$$\mathcal{P}_\vartheta = ((\text{conv}))(\{P_\vartheta | \vartheta \in \boldsymbol{\vartheta}\})$$

"Imprecise Inference" from precise likelihood I

most IP models up to now in statistics:

- precise sampling model $\rightarrow$ likelihood
- imprecise prior about some parameters
- robust Bayesian inference, Bayesian sensitivity analysis (e.g., Rios Insua & Ruggeri (2001, eds. Springer))
- Standard use of GBR in IP, e.g.:
  - Walley (1996, JRSSB)
  - de Cooman & Quaeghebeur (2005, ISIPTA)
  - Walter & Augustin (2009, JStThP)
  - Benavoli & Zaffalon (2014, Statistics)
  - Bickis (2015, ISIPTA)

"Imprecise Inference" from precise likelihood II

- Gamma-maximin decision functions (Noubiap & Seidel (2001, Ann. Stat.))
- Generalized Bayesian updating with other updating rules ("maximum likelihood update"' $\sim$ Dempster rule of conditioning) (Held, Augustin & Kriegler (2008, IJAR))
- interval estimation (confidence regions, credibility regions)
- likelihood regions: traditional; also general framework for likelihood-based decisions (Cattaneo (2007, DissETH Zurich; 2013 EJStat))
- upper probabilities directly derived from the likelihood (Walley & Moral (1999, JRSSB))
- interval-valued logical probability (Weichselberger (2016, in prep))
- Kyburg?, Levi ?

Credal sampling models; imprecise likelihoods

- sets of likelihoods from imprecise data (Plaß et al. (2015, ISIPTA); relation between imprecise data and imprecise sampling models (Schollmeyer (2014, WPMSIIP)))
- imprecise DGP; work so far
  - Walley (Chapter 8.5, 9.6)
  - "likelihood robustness" in robust Bayesianism (Shyamalkumar (2000, in Rios Insua, Ruggeri (eds., Springer)))
  - origin of frequentist robust statistics → neighbourhood models, "capacities instead of probabilities" (Huber (1976, JberDMV)) Huber-Strassen theory: hypotheses tesing, (Huber & Strassen (1973, Ann.Stat.); e.g., Augustin (2002, JSPI))
  - general framework (Hable (2009, DissLMUMunich))
  - minimum distance estimation for single parameter IP model (Hable (2010a, JSPI; 2010b IJAR; imProbEst (CRAN, V1.0))

Terminological Subtleness

- Parametric-Nonparametric
  a) If parameter spaces of infinite dimension are allowed, every nonparametric model is parametric. Similarly, the dimension of parameter spaces should not depend on the sample size
  b) Indeed, every m-dimensional distribution can be formally described by a real-valued parameter[1]
- Precise $\leftrightarrow$ Imprecise sampling models
  Imprecise sampling models may be turned artificially into precise ones: Take an element of the credal set and the deviation from it as a further parameter

[1] Witting (1985, Teubner, p. 5, FN1))

Interpretational Issues $\longrightarrow$ Meaning

- epistemic versus ontological (Walley & Fine (1982)) (disjunctive versus adjunctive) (Weichselberger (2001)), sensitivity analysis point of view versus "?" (Walley (1991))
  - epistemic: true distribution, but only partially known/ described
    - "Likelihood robustness"
    - neighbourhood models in robust statistics
      (type of) true probability distribution, but only partially specified
  - ontological: *set* of probability measures as basic entity
    - Walley (1991, Chapter 9.6)
    - selection rules $\rightarrow$ chaotic probability models (Fierens, Rego, Fine (2009, JSPI), Fierens (2009, IJAR))
    - unobserved heterogeneity, e.g., $\mathbb{E}(Y_i|X_i, Z_i) = f(\beta_0 + \beta_1' X_i + \beta_2 Z_i)$ but Z_i unobservable

Credal (Parametric) Sample Models

- Let $\Theta \subseteq \mathbb{R}$, parametric family of classical distributions $\left(p_{\vartheta|X_i}\right)_{\vartheta\in\Theta}$.
- Credal parametric sampling model (imprecise model, not just imprecise data!).
  Parameter interval-valued

$$\left[\underline{\vartheta}, \overline{\vartheta}\right]$$

  Credal set

$$\mathcal{M}_{X_i} = \left\{P_{\vartheta|X_i} | \vartheta \in \left[\underline{\vartheta}, \overline{\vartheta}\right]\right\}$$

  Strongly independent observations

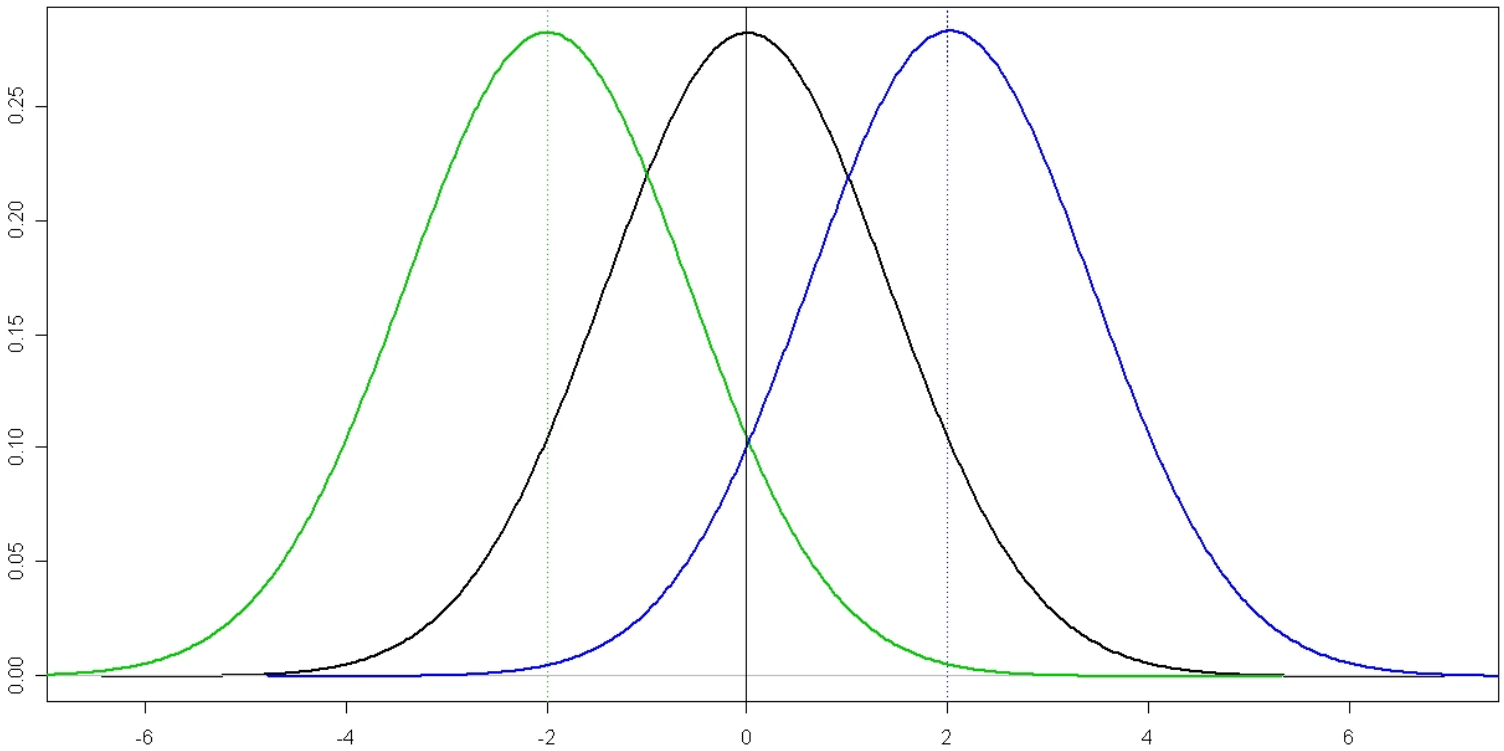
$$\bigtimes_{i=1}^{n} \mathcal{M}_{|X_i} = \left\{\prod_{i=1}^{n} P_{\vartheta_i|X_i} | \vartheta_i \in \left[\underline{\vartheta}, \overline{\vartheta}\right]\right\}$$

What is it good for?

Heterogeneity interpretation:
Overall parameter + individual parameter:

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

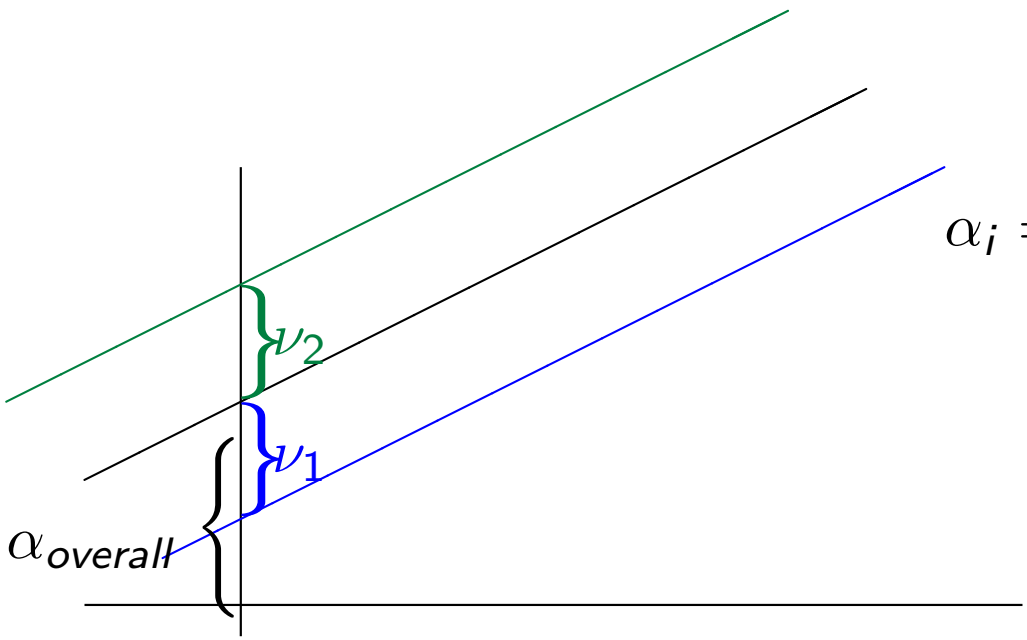
		unobserved
biometrics	overall treatment effect	hospital-, patient-specific
insurance	overall risk	individual risk attitude
dynamical econometric model	overall chance	individual characteristics


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"

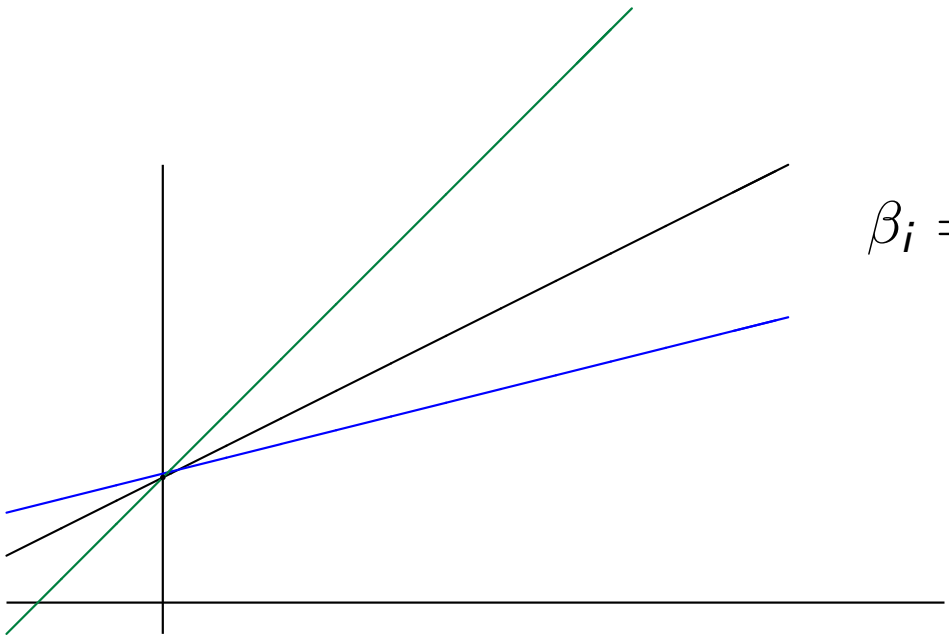
simple linear regression: x_i one-dimensional

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



x_i dummy variables: Analysis of variance with random effects

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



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

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 $\tilde{f}(\cdot)$, claims perfect knowledge of something unobservable

Credal Maximum Likelihood Estimation: A Naive Motivation

- range of parameter values $\hat{\vartheta}_1, \ldots, \hat{\vartheta}_n$ such that

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

- Typical examples: normal distribution leads to

$$\sum_{i=1}^{n} (y_i - \mu_i)^2 \to \min_{\mu_1,\ldots,\mu_n}$$

and

$$\left[\widehat{L\mu}, \widehat{U\mu}\right] = \left[\min_{i=1,\ldots,n} \mu_i, \max_{i=1,\ldots,n} \mu_i\right]$$

$\hat{\mu}_i = y_i$, simply reproducing the sample, no gain in information.

- give more structure to the problem by restricting the variability by constraints on $\left[\widehat{L\vartheta}, \widehat{U\vartheta}\right]$:

$$\widehat{U\vartheta} - \widehat{L\vartheta} \leq \delta$$

for δ given in advance.

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

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

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

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

which can be solved by standard software.

i)

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

ii) The problem can be viewed as a function of the lower interval limit T of the estimator ($\to$ easy calculation)

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

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

Further Aspects/Properties

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 ($\vartheta_i \equiv \vartheta$)

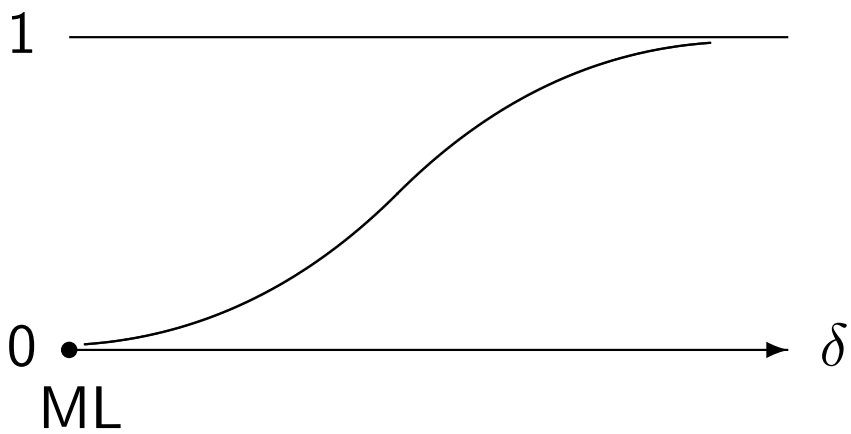
$$\lim_{n\to\infty} P_\vartheta \left(\left[\widehat{L_\delta \vartheta^{(n)}}, \widehat{U_\delta \vartheta^{(n)}} \right] \ni \vartheta \right) = 1$$

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

On the Choice of δ

a) Look at the objective function as a function of δ:
Use that δ for which 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 $\longrightarrow$ credal M-estimators

Conclusion

Claim: In order to utilize the full power of imprecision in statistic, imprecision has to be incorporated into the core of statistical modelling, i.e. into *sampling models*

- Reviewed some aspects of imprecise sampling models and their potential in statistics
- Worked out some "missing links" as research questions
- Briefly discussed one approach $\rightarrow$ credal maximum likelihood, which may be interpreted as an imprecise probability alternative to mixed models

Challenges and Opportunities

- Detailled understanding of statistical properties of credal ML
- Develop imprecise sampling models further
  - Utilize them as a formal superstructure for robustness considerations
  - Utilize different independence concepts for IP
  - Work directly with lower previsions instead of sets of probabilities? But...

"Centuries of probability theory have endowed us with an instinctive intuition about [... sets of probabilities -credal sets] that is hard to compete with. Given that practically everyone grasps the concept of probability – or at least some primitive notion of it – the educational purpose of this framework should clearly not be underestimated. Although I have done my very best to promote various alternatives that have clear advantages, I must admit that at the end of the day, I often think in terms of probabilities. They provide many of the concepts in this dissertation with a valuable intuition and will most likely remain the most important tool for elicitation for a long time to come."
De Bock (2015, Diss U Ghent, p. 292)

Conclusion

Claim: In order to utilize the full power of imprecision in statistic, imprecision has to be incorporated into the core of statistical modelling, i.e. into *sampling models*

- Reviewed some aspects of imprecise sampling models and their potential in statistics
- Worked out some "missing links" as research questions
- Briefly discussed one approach $\rightarrow$ credal maximum likelihood, which may be interpreted as an imprecise probability alternative to mixed models