A Notion of Sufficiency for Statistical Modelling of Interval Data

T. Augustin, E. Endres, M.E.G.V. Cattaneo, P. Fink, J. Plaß, U. Pötter, M. Seitz, G. Schollmeyer, A. Wiencierz

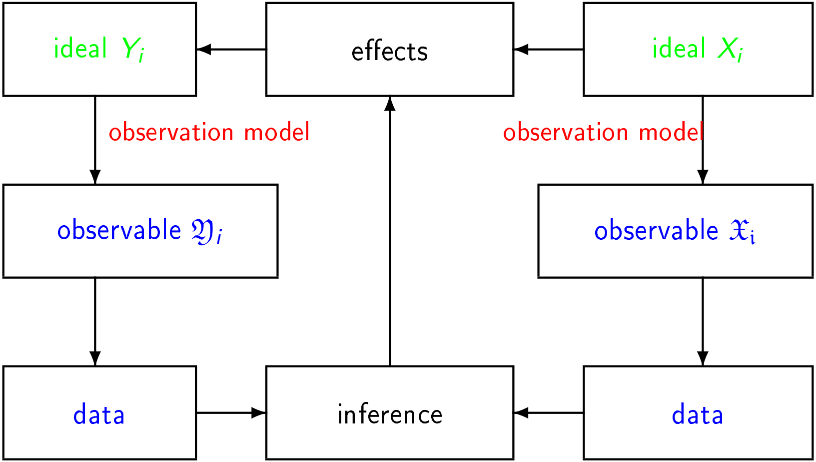
Durham, WPMSIIP 2016

Interval Data

Interval Data

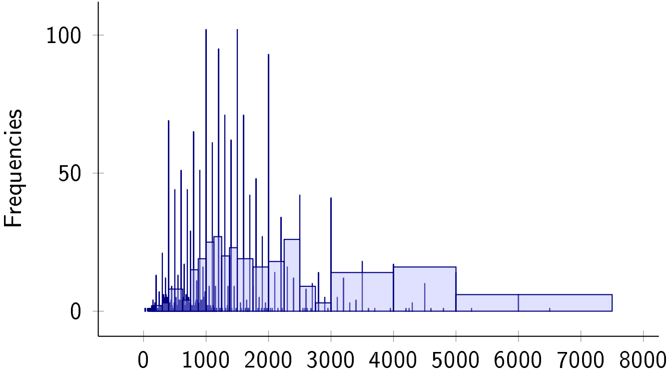
- interval data, more generally "imprecise", "coarse", "messy", "deficient" data are quite common
- *There is an underlying true value that is not observed in the granularity originally intended.*
  epistemic point of view (cp., e.g., Couso & Dubois (2014, IJAR), Couso, Dubois & Sánchez (2014, Springer))
- finite precision of measurements
- response effects like heaping
- anonymization
- compliance, increase of respond rate
- special case: missing data
- categorical data: indecision between certain alternatives
- matching of data
- a better name would be "non-idealized data"

The two-layers perspective



Interval Data: Example

German General Social Survey (ALLBUS) 2010:
2827 observations from Germany in total, 2000 report personal income (30% missing). An additional 10% report only income brackets.



Interval Data: Example

1. We see *heaping* at 1000 €, 2000 €, . . ., less so at 500 €, 1500 €, . . .
2. Both heaping and grouping depend on the amount of income reported.
3. Missingness (some 20% of the data) might as well depend on the amount of income.

Interval Data: Example

1. We see *heaping* at 1000 €, 2000 €, ..., less so at 500 €, 1500 €, ...
2. Both heaping and grouping depend on the amount of income reported.
3. Missingness (some 20% of the data) might as well depend on the amount of income.

Consequences:

1. Missingness, grouping, and heaping will rarely conform to the assumption of "coarsening at random" (CAR).
2. Missingness, grouping, and heaping add an additional type of uncertainty apart from classical statistical uncertainty. This uncertainty can't be decreased by sampling more data.

Interval Data: Example

1. We see *heaping* at 1000 €, 2000 €, ..., less so at 500 €, 1500 €, ...
2. Both heaping and grouping depend on the amount of income reported.
3. Missingness (some 20% of the data) might as well depend on the amount of income.

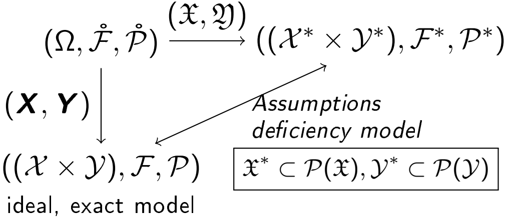
Consequences:

1. Missingness, grouping, and heaping will rarely conform to the assumption of "coarsening at random" (CAR).
2. Missingness, grouping, and heaping add an additional type of uncertainty apart from classical statistical uncertainty. This uncertainty can't be decreased by sampling more data.

Use credible inference procedures that do not rely on unsustainable "assumptions"!

Probability Model

Joint distribution of exact and interval-valued random variables with marginal distributions P (exact data) and P^* (observable, e.g. coarsened data):

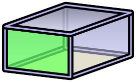


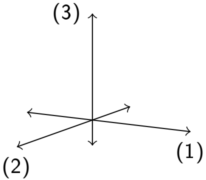
For coarse data: consistency condition (error freeness)

$$\Pr(X \in \mathfrak{X}, Y \in \mathfrak{Y}) = 1$$

Reliable Inference instead of Overprecision

Interval Data: Representations





Epistemic point of view: Couso & Dubois (2014, IJAR), Couso, Dubois & Sánchez (2014, Springer)
We represent interval-valued data as follows:

$$\mathfrak{x} := [\underline{x}, \overline{x}] = \{(x_1, \ldots, x_n) \,|\, \underline{x}_1 \leq x_1 \leq \overline{x}_1, \ldots, \underline{x}_n \leq x_n \leq \overline{x}_n\}$$

where it is assumed that the intervals contain the actual, underlying, "true" $x \in \mathfrak{x}$.
Analogously for Y-variable.

Manski's Law of Decreasing Credibility

Reliability !? Credibility ?

> *"The credibility of inference decreases with the strength of the assumptions maintained." (Manski (2003, p. 1))*

Reliable Inference Instead of Overprecision!!

Consequences from Manski's Law of Decreasing Credibility:

- Adding untenable assumptions to produce precise solution may distroy credibility of statistical analysis, and therefore its relevance for the subject matter questions.
- Make *realistic* assumptions and let the data speak for themselves!
- Extreme case: Consider the *set* of *all* models that are compatible with the data (and then add successively additional assumptions, if desirable)
- The results may be imprecise, but are more reliable
- The extent of imprecision is related to the data quality!
- As a welcome by-product: clarification of the implication of certain assumptions
- Often still sufficient to answer subjective matter question

Work in that direction

- Interval analysis/reliable computing, i.i.d. case, e.g. Nguyen, Kreinovich, Wu, Xiang (2011, Springer)
- Linear regression, e.g.,
  - Rohwer & Pötter (2001, Juventa)
  - Manski & Tamer (2002, Econometrica)
  - Chernozhukov Hong &Tamer (2007, Econometrica)
  - Beresteanu & Molinari (2008, Econometrica)
  - Cattaneo & Wiencierz (2012, IntJAproxReason)
  - Beresteanu, Molchanov,& Molinari. (2012, J Econometrics)
  - Bontemps, Magnac & Maurin (2012, Econometrica)
  - Schollmeyer & Augustin (2015, IntJAproxReason)
- What to do with generalized linear models?
  - logit regression: Plass, Augustin, Cattaneo, Schollmeyer (2015, ISIPTA)
  -
  - Seitz (2015, Springer Best Masters)

Generalized Linear Models; Maximum Likelihood Estimation

Basic Notation, Regression Models

- n observations („large “)
- $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^T$ response variable
- $\boldsymbol{X} = (X_1, \cdots, X_n)^T$ covariates
- $(X_i, Y_i)_{i=1,\cdots,n}$ i.i.d
- here Y_i one dimensional, of metrical, ordinal, or categorical scale
- X_i p-dimensional, (metric or binary)
- joint distribution: density with respect to appropriate dominating measure

$$f_{(\boldsymbol{X},\boldsymbol{Y})}(\boldsymbol{x},\boldsymbol{y}) = \prod_{i=1}^{n} f_{(X_i,Y_i)}(x_i, y_i) = \prod_{i=1}^{n} \underbrace{f_{Y_i|X_i}(y_i|x_i)}_{model} \cdot f_{X_i}(x_i)$$

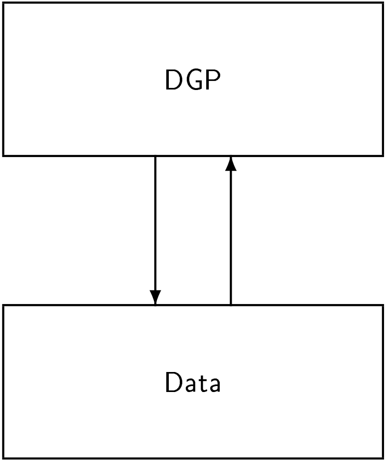
- Typically parametrization of $f_{Y|X}(\cdot)$ only,
  $f_X(\cdot)$ is assumed to contain ancillary information
- regression parameters $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$, further parameter γ
- parametric model for $[Y_i|X_i]$
- Here generalized linear model

- E.g. Fahrmeir, Kneib, Lang, Marx (2013, Spinger)
- Generalizing linear regression

$$Y_i = \beta_0 + \beta_1' X_i + \varepsilon_i \iff Y_i | X_i \sim N(X_i' \beta, \sigma^2)$$

  to other distributions
  * Gamma distribution, inverted Gaussian, Beta distribution
  * Poisson distribution $\longrightarrow$ count data
  * Bernoulli/Multinomial distribution $\longrightarrow$ categorical data: logit/Probit model
- $f(y_i || \nu_i, \gamma) = \text{const}(y_i, \gamma) \cdot \exp(\frac{\nu_i y_i - b(\vartheta_i)}{\gamma}), \ i = 1, \cdots, n$
- $\nu_i = \beta_0 + \beta_1 \cdot x_{i1} + \cdots + \beta_p \cdot x_{ip}$
- exponential family with individual canonical parameter $\nu_i = \begin{pmatrix} 1 \\ X_i' \end{pmatrix}' \beta$ ("canonical link")

DGP



Maximum Likelihood Estimation

- After having observed the data, reinterpret the density as a function of the parameters, describing how likely each parameter has produced the data.
- Maximum Likelihood-Estimator (MLE): root of the derivative of the logarithmized likelihood $\longrightarrow$ score function

$$\mathrm{score}(\beta) = \frac{1}{\gamma} \sum_{i=1}^{n} \begin{pmatrix} 1 \\ X_i \end{pmatrix} (Y_i - \mathbb{E}(Y_i|X_i))$$

- For discussion later; general form

$$\mathrm{score}(\beta) = \boldsymbol{X}\boldsymbol{D}(\beta)\sigma^2(\beta) \cdot (\boldsymbol{Y} - \mathbb{E}(Y_i|X_i)$$

- Quasi-likelihood models
- multivariate Y
- "Weibull-type": Y_i^α, $Y_i \geq 0$

$\mathbb{E}(Y_i|X_i) = h(\eta_i)$ response function
and
$g(\mathbb{E}(Y_i|X_i)) = \eta_i$ link function
$\mathbb{E}(Y_i|X_i) = b'(\vartheta_i)$, $\vartheta_i = \psi(\mathbb{E}(Y_i|X_i))$
$Var(Y_i|X_i) = \phi \cdots$

Collecting Regions from Estimating Equations

Estimating Equations–> Collection Regions

Generalizing from the linear case, suppose there is a consistent (score-) estimating equation for the ideal model $\{\mathcal{P}_\vartheta \,|\, \vartheta \in \Theta\}$, i.e.:

$$\forall \vartheta \in \Theta : \quad \mathbb{E}_\vartheta \left(\psi(\boldsymbol{X}, \boldsymbol{Y}; \vartheta) \right) = 0$$

Then

$$\hat{\vartheta} := \mathrm{root} \left(\psi(\boldsymbol{X}, \boldsymbol{Y}; \vartheta) \right)$$

With interval data, one gets a set of estimating equations, one for each random vector (selection) $(\boldsymbol{X}, \boldsymbol{Y}) \in (\mathfrak{X}, \mathfrak{Y})$:

$$\Psi(\mathfrak{X}, \mathfrak{Y}; \vartheta) := \{\Psi(\boldsymbol{X}, \boldsymbol{Y}; \vartheta) \,|\, \boldsymbol{X} \in \mathfrak{X}, \boldsymbol{Y} \in \mathfrak{Y}\}$$

$$\hat{\Theta} := \left\{ \hat{\vartheta} \,\middle|\, \exists \boldsymbol{X} \in \mathfrak{X}, \boldsymbol{Y} \in \mathfrak{Y} \, : \, \hat{\vartheta} = \mathrm{root} \left(\psi(\boldsymbol{X}, \boldsymbol{Y}; \vartheta) \right) \right\}$$

Named “collection region” in Schollmeyer & Augustin (2015, IntJAproxReason)

Envelopes of Estimating Equations: One Dimensional Case

Envelopes of Estimating Equations: One Dimensional Case

Seitz (2015, Springer Best Masters, § 3.1)

- Common form of estimating function

$$\psi(X, Y; \vartheta) = \sum_{i=1}^{n} \psi_i(X_i, Y_i; \vartheta).$$

- ϑ one-dimensional then

$$\min_{(X,Y)\in(\mathfrak{X},\mathfrak{Y})} \psi(X, Y; \vartheta) = \sum_{i=1}^{n} \min_{(X,Y)\in(\mathfrak{X},\mathfrak{Y})} \psi_i(X_i, Y_i, \vartheta)$$

If sign of derivative of the score function does not change, Fisher scoring; based on the sum of the individual lower and upper envelopes of the score functions, which usually can be calculated analytically

One Parameter Case

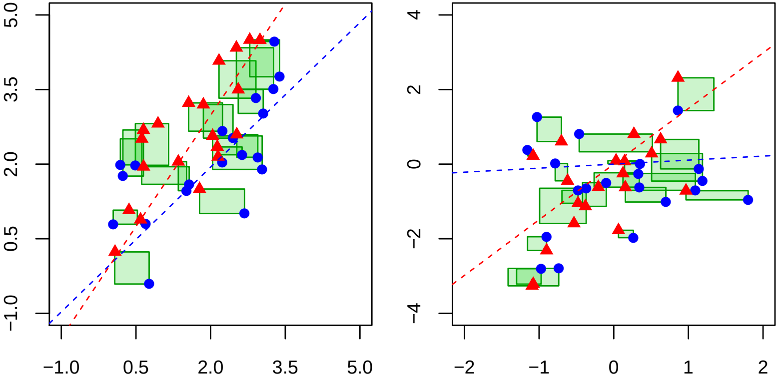


Figure: Simulation; linear model without intercept.

Exponential

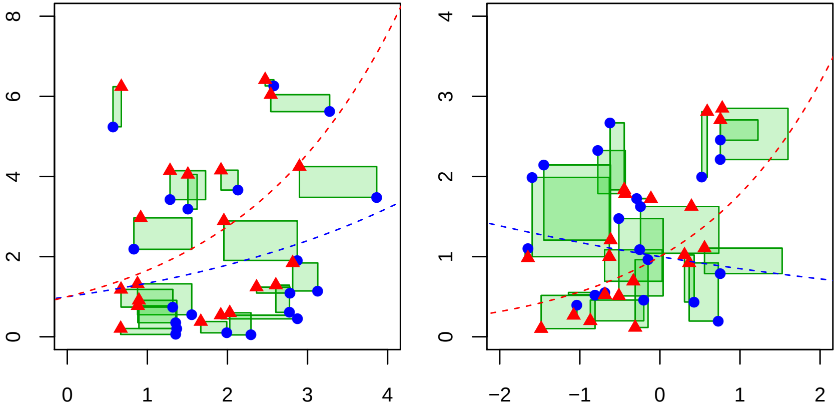


Figure: Exponential case

Penalty Approach

Parameter Estimation, Basic Form

- Linear objective function with nonlinear equality constraint and box constraints:

$$\vartheta_l \to \text{min / max}$$

subject to

$$\begin{array}{rcll} \psi_k(x, y; \vartheta) & = & 0 & \text{with} \quad k = 1, \ldots, q \\ x_i & \in & \mathfrak{X}_i & \text{with} \quad i = 1, \ldots, n \\ y_i & \in & \mathfrak{Y}_i & \text{with} \quad i = 1, \ldots, n. \end{array}$$

Parameter Estimation, Penalty Form

Seitz (2015, Springer Best Masters, § 3.5, 4)

- $\hat{\vartheta}$ root of function $\psi(\cdot) \iff \hat{\vartheta} := \mathrm{argmin}_{\vartheta}\,(\psi)^2$
- Nonlinear objective function with box constraints:

$$\vartheta_l \pm \sum_{k=1}^{q} \rho_k \left(\psi_k(x, y; \vartheta)\right)^2 \to \min \,/\, \max$$

subject to $x \in \mathfrak{x}, \quad y \in \mathfrak{y}$
$\rho_k, k = 1, \ldots, q$ penalties

Parameter Estimation: Heuristic Search

Sequential evaluation

- Fix X, Y
- Search for optimal vertex in $(\mathfrak{X}_1 \times \mathfrak{Y}_1)$
- Fix this optimum and search for optimal vertex in $(\mathfrak{X}_2 \times \mathfrak{Y}_2)$ etc.
- Repeat until no considerable change in optimal solution

MLE-Equivalence

Def: MLE-equivalence for $A\theta$

Let $\mathcal{P}$ be a family of distributions parametrized in $\vartheta \in \Theta \subseteq \mathbb{R}^q$ and denote for each sample $(\boldsymbol{X}, \boldsymbol{Y}) \sim p_\vartheta \in \mathcal{P}$ the maximum likelihood estimator for ϑ by $\hat{\vartheta}(\boldsymbol{X}, \boldsymbol{Y})$.
For a matrix $A \in \mathbb{R}^{\tilde{q} \times q}$, $\tilde{q} \leq q$ call two samples $(\boldsymbol{X}^{(1)}, \boldsymbol{Y}^{(1)})$ and $(\boldsymbol{X}^{(2)}, \boldsymbol{Y}^{(2)})$ *MLE-equivalent for $A\theta$* if

$$A\hat{\vartheta}\left(\boldsymbol{X}^{(1)}, \boldsymbol{Y}^{(1)}\right) = A\hat{\vartheta}\left(\boldsymbol{X}^{(2)}, \boldsymbol{Y}^{(2)}\right)$$

Examples

- For arbitrary A and sample $(\boldsymbol{X}, \boldsymbol{Y})$, let $\left(\boldsymbol{X}^{(1)}, \boldsymbol{Y}^{(1)}\right) = (\boldsymbol{X}, \boldsymbol{Y})$ and $\left(\boldsymbol{X}^{(2)}, \boldsymbol{Y}^{(2)}\right)$ be an order statistic of $(\boldsymbol{X}, \boldsymbol{Y})$ with respect to one of its components
- Of particular interest are specific A's such that certain subvectors of components of $\vartheta = (\beta^T, \zeta^T)^T$ are selected, in particular A such that $A\vartheta = \beta$
  $\Rightarrow$ MLE-equivalent for β

Theorem

GLM with canonical link functions and $\boldsymbol{X}$ treated as fixed
all $\left(\boldsymbol{X}^{(1)}, \boldsymbol{Y}^{(1)}\right)$ and $\left(\boldsymbol{X}^{(2)}, \boldsymbol{Y}^{(2)}\right)$ with

$$\sum_{i=1}^{n} \begin{pmatrix} 1 \\ X_{i1}^{(1)} \\ \vdots \\ X_{ip}^{(1)} \end{pmatrix} \cdot Y_i^{(1)} = \sum_{i=1}^{n} \begin{pmatrix} 1 \\ X_{i1}^{(2)} \\ \vdots \\ X_{ip}^{(2)} \end{pmatrix} \cdot Y_i^{(2)}$$

are MLE-equivalent for β.

For the proof remember:

MLE for β from the score function

$$\mathrm{score}(\beta) = \frac{1}{\gamma} \sum_{i=1}^{n} \begin{pmatrix} 1 \\ X_i \end{pmatrix} (Y_i - \mathbb{E}(Y_i | X_i))$$

Corollary

To calculate the collection region for fixed covariates and interval valued response it suffices to consider certain single representers of MLE equivalent samples.

Algorithm ($\boldsymbol{X}$ precise)

Instead of solving the nonlinear (even nonconvex!) optimization problem in the penalty approach with n box constraints, determine the p-dimensional "variational areatt" of

$$\sum_{i=1}^{n} \begin{pmatrix} 1 \\ X_{i1} \\ \vdots \\ X_{ip} \end{pmatrix} \cdot Y_i .$$

This is linear and even can be described explicitly. ((One dimensional X, w.l.o.g. $X > 0$: Sort by X: Start with taking all minimal Y's. The next point is as large (small) as possible by using that unit with the highest (the smallest) X value and the corresponding Y_{max} (Y_{min}).))
Then work with representers from there.

Lemma

If domain of covariates is compact, then, without loss of generality, all covariates can be taken to be positive

for one dimension

$$min\, X := min_{i=1,\dots,n} X_i > 0$$

else consider

$$X_i^+ := X_i - min\, X > 0$$

regression with

$$\beta_0^+ + \beta_1^+ X_i = \beta_0^+ + \beta^+ X_i - \beta^+ min\, X = \tilde{\beta}_0 + \beta^+ X_i$$

Corollary

Consider only regression model with a linear predictor and regression parameter $(\beta_0, \beta_1, \ldots, \beta_p)'$:

$$(\tilde{X}_i, Y_i)_{i=1,\ldots,n} \text{ and } (X_i, Y_i)_{i=1,\ldots,n},$$

where

$$\tilde{X}_i = X_i + c, \; c \in \mathbb{R},$$

are MLE-equivalent for $(\beta_1, \ldots, \beta_p)'$.

In more detail

Let $\boldsymbol{X}$ be one dimensional.
Consider for
$X = (X_1, ..., X_n)$
the order statistics
$\boldsymbol{X} \uparrow := (X_{(1)}, \ldots, X_{(n)})$
and the reverse order statistics
$\boldsymbol{X} \downarrow := (X_{(n)}, \ldots, X_{(1)})$

Sort $\underline{Y}$ and $\overline{Y}$ accordingly

$$\begin{aligned} \underline{\boldsymbol{Y}}\uparrow^{\boldsymbol{x}} &= (\underline{Y}_{[1]}, \underline{Y}_{[2]}, \ldots, \underline{Y}_{[n]}) \\ \overline{\boldsymbol{Y}}\uparrow^{\boldsymbol{x}} &= (\overline{Y}_{[1]}, \overline{Y}_{[2]}, \ldots, \overline{Y}_{[n]}) \end{aligned}$$

Describe vertices of "upper polygon"', starting from

$$\left(\sum_{i=1}^{n} \underline{Y}_i \, , \, \sum_{i=1}^{n} \underline{Y}_i X_i \right)$$

order statistics:

$$\begin{array}{rcl} \boldsymbol{X} & = & (X_{(1)}, \ldots, X_{(n)}) \\ \text{sort} & \underline{\boldsymbol{Y}}, \overline{\boldsymbol{Y}} & \text{accordingly} \\ \underline{\boldsymbol{Y}} \uparrow^{x} & = & (\underline{Y}_{[1]}, \underline{Y}_{[2]}, \ldots, \underline{Y}_{[n]}), \text{ i.e.} \\ \underline{\boldsymbol{Y}} \downarrow_{x} & = & (Y_{[n]}, Y_{[n-1]}, \ldots, Y_{[1]}) \end{array}$$

etc.

first vertex further on:

- increase $\sum_{i=1}^{n} \underline{Y}_i$ by ϵ
- highest (lowest) point i
  put all mass into the largest (smallest) $\boldsymbol{X}$-value

vertices of lower envelope ($\sum_{\phi} := 0$)

$$\left(\sum_{i=1}^{j} \overline{Y}_{[i]} + \sum_{i=j+1}^{n} \underline{Y}_{[i]}, \sum_{i=1}^{j} \overline{Y}_{[i]} X_{(i)} + \sum_{i=j+1}^{n} \underline{Y}_{[i]} X_{(i)} \right)$$

vertices of upper envelope

$$\left(\sum_{i=1}^{j} \overline{Y}_{[n+1-i]} + \sum_{i=j+1}^{n} \underline{Y}_{[n+1-i]}, \sum_{i=1}^{j} \overline{Y}_{[n+1-i]} \cdot X_{(n+1-i)} + \sum_{i=j+1}^{n} \underline{Y}_{[n+1-i]} \cdot X_{(n$$

Explicit characterization of vertices.

$\Rightarrow$ check for given $\vec{\beta}^*$ whether or or not it is in the collection region.

Concluding Remarks

Concluding Remarks

- Interval (coarse(ned)) data in generalized linear models
- Optimization approach based on score function
- Try to make it more tractable by „MLE-equivalence “
- $\Rightarrow$ Sufficiency concept for coarse data (interval data)