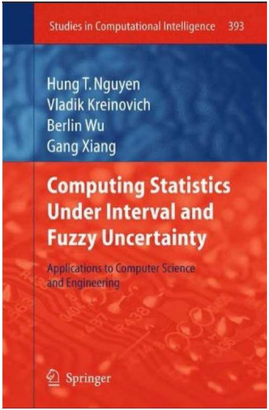
Some applications of interval computing in statistics

Michal Černý

Department of Econometrics & DYME Research Center
University of Economics, Prague, Czech Republic

SWIM 2015, Prague

- Many ideas and results are summarized in the wonderful book:



- Some results: joint research with M. Hladík, M. Rada, O. Sokol, J. Horáček, J. Antoch et al.

The core problem of Interval Analysis

- We are given a (continuous, say) function $f : \mathbb{R}^n \to \mathbb{R}$ and a box $\mathbf{x} \in \mathbb{IR}^n$. We are to determine the range

$$f(\mathbf{x}) = [\underline{f(\mathbf{x})}, \overline{f(\mathbf{x})}] = \{f(x) : x \in \mathbf{x}\}.$$

- Which particular functions f are interesting in statistics & data analysis?

- **Outline:**
  - Part I: one-dimensional interval-valued data
  - Part II: multivariate data & regression

Part I. One-dimensional data

One-dimensional data: a model

Assumptions.

- Let $x_1, \ldots, x_n$ be a dataset; for example, let the data be a random sample from a distribution Φ. The dataset is unobservable.
- What is observable is a collection of intervals $\mathbf{x}_1, \ldots, \mathbf{x}_n$ such that

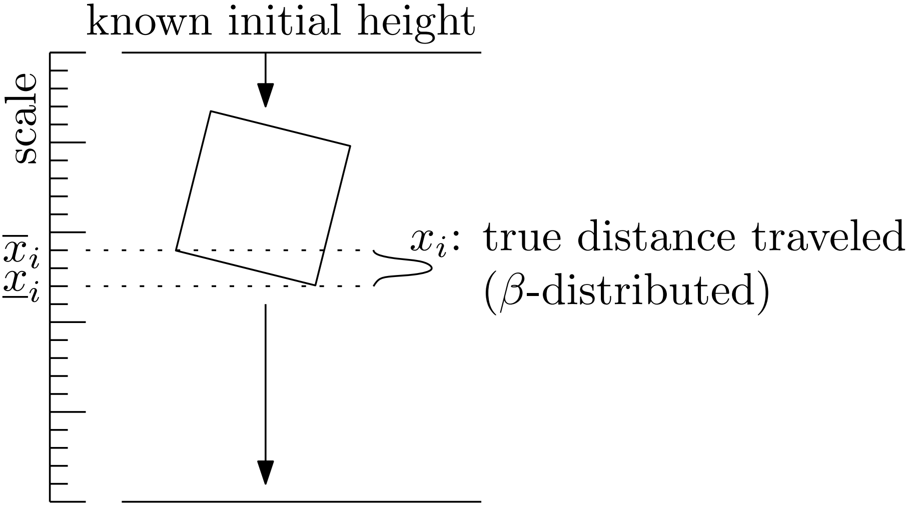
$$x_1 \in \mathbf{x}_1, \ldots, x_n \in \mathbf{x}_n \quad \text{a.s.}$$

- **A general goal:** We want to make inference about the original dataset $x_1, \ldots, x_n$, about the generating distribution Φ, about its parameters, we want to test hypotheses etc.
- We are given a statistic $S(x_1, \ldots, x_n)$ and we want to determine/estimate its value, distribution, or other properties, using only the observable interval-valued data $\mathbf{x}_1, \ldots, \mathbf{x}_n$.
- **Now:** the appropriate toolbox depends on whether we can make further assumptions on the distribution of $(x, \mathbf{x})$.

Example

- Allmaras et al., SIAM Review 55 (2013); Aguilar et al., SIAM Review 57 (2015)
- Measurement of a **falling box**: the aim is to estimate the gravity acceleration and air resistance
- A camera takes snaps in discrete times: the position x_i ($=$ distance traveled in time i) is uncertain due to unpredictable rotation
- They make an assumption that the distribution of x_i given $\underline{x}_i, \overline{x}_i$ is beta and apply Bayesian framework

Example (contd.)



The possibilistic approach

- Interval computation comes into play when the only assumption about the distribution of $(x, \mathbf{x})$ we make is $x \in \mathbf{x}$ a.s. Nothing more.
- Then, given a statistic S, the only information we can infer about S from the observable interval-valued data $\mathbf{x}$ is the pair of tight bounds

$$\overline{S} = \max\{S(\xi) : \xi \in \mathbf{x}\},$$
$$\underline{S} = \min\{S(\xi) : \xi \in \mathbf{x}\},$$

clearly satisfying

$$\underline{S} \leqslant S(x) \leqslant \overline{S} \quad \text{a.s.}$$

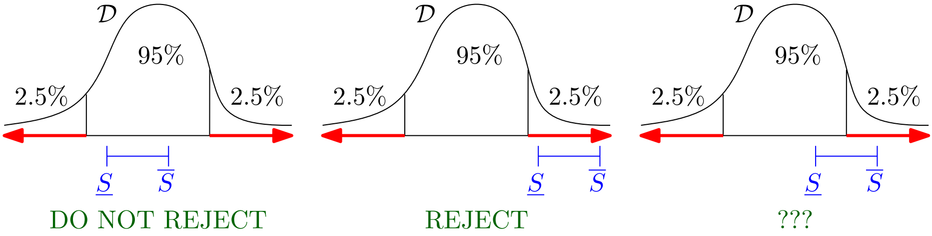
- **Remark.** In econometrics, partial knowledge about the distribution $(x, \mathbf{x})$ is referred to as partial identification: see the survey paper E. Tamer, **Partial identification in econometrics**, Annual Review of Economics 2 (2010), pp. 167–195.
- Also many papers in Econometrica and other journals.

Which statistics are interesting?

- Descriptive statistics: sample mean, variance, median, coefficient of variation, quantiles, higher-order moments, . . .
- Many well-known people did a lot of work: *Kreinovich, Ferson, Ginzburg, Aviles, Longpré, Xiang, Ceberio, Dantsin, Wolpert, Hajagos, Oberkampf, Jaulin, Patangay, Starks, Beck, . . .* (sorry that I cannot mention all)
- Estimators of parameters of the data-generating distribution Φ
- Test statistics for various hypotheses

Test statistics

- We are to test a null hypothesis (H_0) against an alternative A
- We usually construct a test statistic S s.t. its distribution $\mathcal{D}$ under H_0 is known
- Then, quantiles of $\mathcal{D}$ determine the critical region, where we reject H_0 at a pre-selected level α of confidence (say, $\alpha = 95\%$)
- Given the intervals $\mathbf{x}_1, \ldots, \mathbf{x}_n$: if we can compute $\underline{S}, \overline{S}$, then we can make at least partial conclusions:



Test statistics: An example

- **Example.** Say that $x_1, \ldots, x_{n/2}$ and $x_{(n/2)+1}, \ldots, x_n$ are two independent samples from $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively.
- We want to test stability of variance: $\sigma_1^2 = \sigma_2^2$.
- A well-known test statistic: F-ratio

$$F = \frac{\text{sample variance of } x_1, \ldots, x_{n/2}}{\text{sample variance of } x_{(n/2)+1}, \ldots, x_n}.$$

- Problem: computation of both values $\underline{F}, \overline{F}$ is NP-hard! (How serious is this obstacle? We will see later...)

- Let $x_1, \dots, x_n$ be a $N(\mu, \sigma^2)$ sample
- Given $\mu_0 \in \mathbb{R}$, to test $\mu = \mu_0$ we use the t-ratio (coefficient of variation)

$$t = \frac{|\widehat{\mu} - \mu_0|}{\widehat{\sigma}} = \frac{\left|\left(\frac{1}{n}\sum_{i=1}^n x_i\right) - \mu_0\right|}{\sqrt{\frac{1}{n-1}\sum_{j=1}^n \left(x_j - \frac{1}{n}\sum_{k=1}^n x_k\right)^2}}.$$

- **Some results:**
  - $\underline{t}$ is NP-hard and inapproximable with an arbitrary absolute error
  - $\underline{t}$ is computable in pseudopolynomial time
  - $\overline{t}$ computable in polynomial time

Test statistics: Further examples

- Testing independence: Durbin-Watson statistic

$$DW = \frac{\sum_{i=2}^{n}(r_i - r_{i-1})^2}{\sum_{j=1}^{n} r_i^2},$$

  where $r_i = x_i - \frac{1}{n}\sum_{k=1}^{n} x_k$.
- Testing stability of mean (important e.g. in quality control):
  - H_0: $\mathrm{E}x_1 = \mathrm{E}x_2 = \cdots = \mathrm{E}x_n$
  - A:
    $\exists k: \ \mathrm{E}x_1 = \mathrm{E}x_2 = \cdots = \mathrm{E}x_k = \mu_1 \neq \mu_2 = \mathrm{E}x_{k+1} = \mathrm{E}x_{k+2} = \cdots = \mathrm{E}x_n.$
  - Test statistic:

$$T = \max_{k=1,\ldots,n-1} \frac{\sqrt{\frac{n}{k(n-k)}}\sum_{\ell=1}^{k}(x_\ell - \frac{1}{n}\sum_{\iota=1}^{n} x_\iota)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n} x_j\right)^2}}.$$

- Computational aspects of $\underline{S}$ and $\overline{S}$ have been investigated for many statistics S... and many are still waiting...

Sample variance

$$\overline{s^2} = \max\left\{ \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \frac{1}{n} \sum_{j=1}^{n} x_j \right)^2 \ : \ x \in \mathbf{x} \right\},$$

$$\underline{s^2} = \min\left\{ \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \frac{1}{n} \sum_{j=1}^{n} x_j \right)^2 \ : \ x \in \mathbf{x} \right\}.$$

- **Observation:** $\underline{s^2} \to$ CQP $\to$ weakly polynomial time
- **Ferson et al.:** a strongly polynomial algorithm $O(n^2)$
- **Unfortunately:** $\overline{s^2}$ is NP-hard
- **Even worse:** $\overline{s^2}$ is NP-hard to approximate with an arbitrary absolute error

NP-hardness of $\overline{s^2}$

- NP-hardness of $\overline{s^2}$ $\to$ investigation of special cases solvable in polynomial time
- **Ferson et al.**: consider the "$\frac{1}{n}$-narrowed" intervals

$$\tfrac{1}{n}\mathbf{x}_i := [x_i^C - \tfrac{1}{n}x_i^\Delta, x_i^C + \tfrac{1}{n}x_i^\Delta], \quad i = 1, \ldots, n.$$

Theorem: If $\frac{1}{n}\mathbf{x}_i \cap \frac{1}{n}\mathbf{x}_j = \emptyset$ for all $i \neq j$, then $\overline{s^2}$ can be computed in polynomial time.

- Another formulation: If there is no k-tuple of indices $1 \leqslant i_1 < \cdots < i_k \leqslant n$ such that

$$\bigcap_{\ell \in \{i_1, \ldots, i_k\}} \tfrac{1}{n}\mathbf{x}_\ell \neq \emptyset,$$

then $\overline{s^2}$ can be computed in time $O(p(n) \cdot 2^k)$, where p is a polynomial.

Computation of $\overline{s^2}$ & Ferson et al. (contd.)

Graph-theoretic reformulation: Let $G_n(V_n, E_n)$ be the interval graph over $\frac{1}{n}\mathbf{x}_1, \ldots, \frac{1}{n}\mathbf{x}_n$:

- Vertices: $V_n =$ set of the narrowed intervals $\frac{1}{n}\mathbf{x}_1, \ldots, \frac{1}{n}\mathbf{x}_n$
- Edges: $\{i, j\} \in E$ $(i \neq j)$ iff $\frac{1}{n}\mathbf{x}_i \cap \frac{1}{n}\mathbf{x}_j \neq \emptyset$
- Let ω_n be the size of the largest clique of G_n. Now: the algorithm works in time $O(p(n) \cdot 2^{\omega_n})$.

Remark. Determining the largest clique of an interval graph is easy.

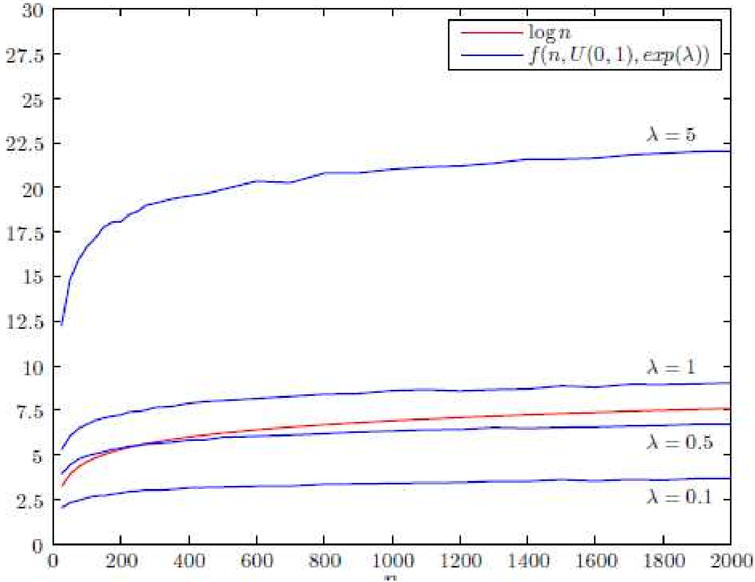
Remark. The worst case is bad — e.g. when $\mathbf{x}_1^C = \mathbf{x}_2^C = \cdots = \mathbf{x}_n^C$. (Such instances result from the NP-hardness proof.)

But: What if the data are generated by a random process? Then, do the "ugly" instances occur frequently, or only rarely?

Simulations

Assumption: The centers and radii of intervals $\mathbf{x}_i$ are generated by a "reasonable" random process:

- Centers $\mathbf{x}_i^C$: sampled from a "reasonable" distribution (continuous, finite variance) — uniform, normal, exp, ...
- Radii $\mathbf{x}_i^\Delta$: sampled from a "reasonable" nonnegative distribution (continuous, finite variance) — uniform, one-sided normal, exp, ...
- Simulations show Sokol's conjecture: The clique is logarithmic on average! Thus: The algorithm is polynomial on average.



30
27.5
25
22.5
20
17.5
15
12.5
10
7.5
5
2.5
0
0
200
400
600
800
1000
1200
1400
1600
1800
2000
$\log n$
$f(n, U(0,1), exp(\lambda))$
$\lambda = 5$
$\lambda = 1$
$\lambda = 0.5$
$\lambda = 0.1$

Sokol's conjecture II

Furthermore: It seems that $\mathrm{var}(\omega_n) = O(1)$ ("Sokol's conjecture II').

- Say, for simplicity, that indeed $\mathrm{E}\omega_n = \log n$. By Chebyshev's inequality we get:

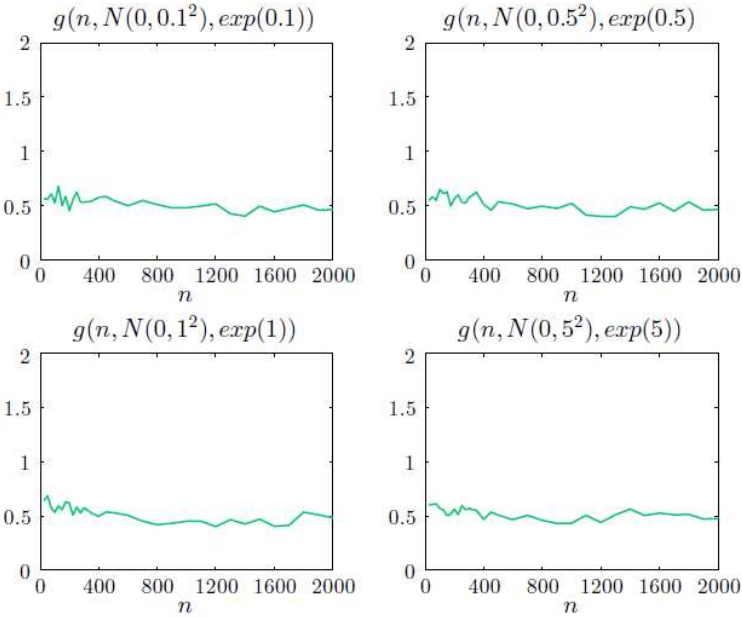
$$\Pr[\omega_n \geqslant \log n + \underbrace{10\sqrt{\mathrm{var}(\omega_n)}}_{=:K \text{ (constant)}}] \leqslant 1\%.$$

- Thus: in 99% cases, the algorithm of Ferson et al. works in time at most

$$p(n) \cdot 2^{K + \log n},$$

where K does not grow with n.

Sokol's conjecture II



A pair of remarks

To summarize:

- We have random intersection (interval) graphs and we need to estimate the average size of the largest clique and its variance
- This department has a strong tradition both in intersection graphs and random graphs — Jiří Matoušek (†2015); Jan Kratochvíl et al.
- Interesting problem: our model of a random graph is different from the traditional models $G_{n,p}$ and $G_{n,m}$

Computation of $\overline{s^2}$: Xiang et al.

Another interesting algorithm by Xiang et al.:

- **Definition.** If there is no pair of indices i, j such that

$$\tfrac{1}{n}\mathbf{x}_i \subseteq \text{interior}(\tfrac{1}{n}\mathbf{x}_j),$$

  we say that the dataset $\mathbf{x}_1, \ldots, \mathbf{x}_n$ satisfies the no-subset property.

- **Remark.** Very natural when the intervals have the same radii — e.g. when the data have been measured by the same device with a single error radius.
- **Theorem.** If the dataset satisfies the no-subset property, then $\overline{s^2}$ can be computed in polynomial time.
- **A more general statement:** Let $J \subseteq \{1, \ldots, n\}$ be a set of indices such that the dataset $\{\mathbf{x}_i : i \in \{1, \ldots, n\} \setminus J\}$ satisfies the no-subset property. Then $\overline{s^2}$ can be computed in time $O(p(n) \cdot 2^{|J|})$.

A brief summary

- Further good news: $\overline{s^2}$ is computable pseudopolynomially
- Main message: although NP-hard in theory, $\overline{s^2}$ is efficiently computable "almost always" (in the probabilistic setup) — hard instances are rare
- A nice interdisciplinary problem: statistical motivation, interval-theoretic and graph-theoretic methods

A pair of remarks

- (Some) ideas can be (sometimes) generalized: observe that s^2 can be written as

$$s^2 = \frac{1}{n-1} Q^2 - \frac{1}{n(n-1)} L^2,$$

where $Q^2 = \sum_i x_i^2$ and $L = \sum_i x_i$.

- Many more statistics can be written as "simple functions" of Q and L, e.g. the t-ratio (coefficient of variation):

$$t^2 = \frac{\frac{1}{n^2} L^2}{\frac{1}{n-1} Q^2 - \frac{1}{n(n-1)} L^2}.$$

- Recall also the F-ratio

$$F = \frac{\text{sample variance of } x_1, \ldots, x_{n/2}}{\text{sample variance of } x_{(n/2)+1}, \ldots, x_n};$$

positive results for sample variance apply here directly, too.

Part II. The multivariate case

The core problem of Interval Analysis more generally

- We are given a (continuous, say) function $f : \mathbb{R}^n \to \mathbb{R}^m$ and a box $\mathbf{x} \in \mathbb{IR}^n$.
- We are to say something reasonable about the range

$$f(\mathbf{x}) = \{f(x) \in \mathbb{R}^m : x \in \mathbf{x}\}.$$

Motivation: Joint regions for dependent statistics

- Statistics are often dependent: we are, e.g., interested in the joint region for

$$S(x_1, \ldots, x_n) = (\text{sample mean}, \text{sample variance}) \in \mathbb{R}^2, \quad x \in \mathbf{x}.$$

- J. Stoye, Partial identification of spread parameters, Quantitative Economics, 2010: “ This paper analyzes partial identification of parameters that measure a distribution’s spread, for example, the variance, Gini coefficient, entropy, or interquartile range. The core results are tight, two-dimensional identification regions (that are typically not rectangles) for the expectation and variance, the median and interquartile ratio, and many other combinations of parameters. They are developed for numerous identification settings, including but not limited to cases where one can bound either the relevant cumulative distribution function or the relevant probability measure. Applications include missing data, interval data, (...) contaminated data (...). “

J. Stoye (Quant Econ, 2010): Example

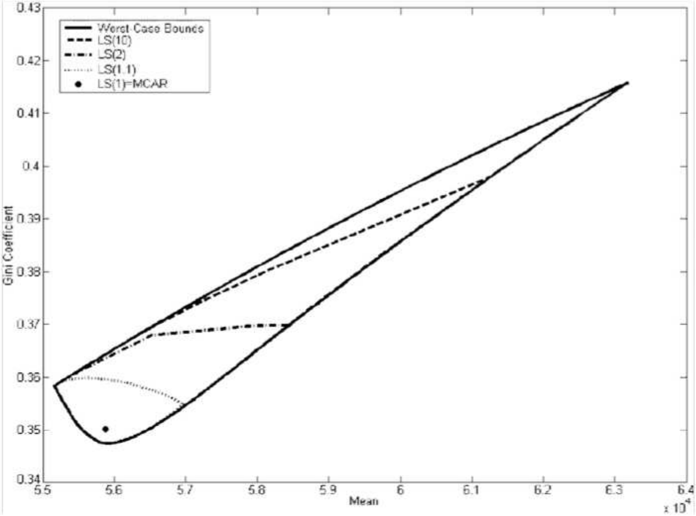


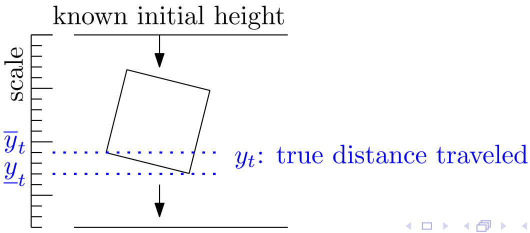
FIGURE 4. Joint identification of mean and Gini coefficient (CPS data).

Regression

- The most important statistical application (J. Á. Víšek: "95% of practical statistical problems involve regression")
- The most practically important joint region of dependent statistics: the set of estimates of regression coefficients
- Recall the falling box example: regression model

$$y_t = \frac{1}{C} \log \cosh \left(\sqrt{gC}(t - t_0) \right) + \varepsilon_t,$$

following from Newton's equations, where C, g, t_0 are unknown parameters, ε_t is random noise, and the dependent variable y_t is the (interval-valued) distance traveled.



Regression

- A general form of the linear regression model with interval data:

$$y = X\theta + \varepsilon, \quad y \in \mathbf{y},\ X \in \mathbf{X},$$

  where observable data are $(\mathbf{X}, \mathbf{y})$ and the only property of the joint distribution of $(X, \mathbf{X}, \varepsilon, \mathbf{y})$ is that $y \in \mathbf{y},\ X \in \mathbf{X}$ holds a.s.
- The most important statistics:
  - $\widehat{\theta} = \widehat{\theta}(X, y)$ $(\in \mathbb{R}^p)$: an estimator
  - $R = R(X, y) = \|y - X\widehat{\theta}\|$ $(\in \mathbb{R})$: loss function (goodness-of-fit measure); here $\|\cdot\|$ is some vector norm
- A nice case (observed by Schön, Kutterer and others): if $\underline{X} = \overline{X} =: X$ and $\widehat{\theta}$ is the least-squares estimator, then the joint region of estimates

$$\{\theta^* \in \mathbb{R}^p : \theta^* = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y,\ y \in \mathbf{y}\}$$

  is a zonotope in the parameter space.
- The general case $\{\theta^* \in \mathbb{R}^p : \theta^* = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y,\ y \in \mathbf{y}, X \in \mathbf{X}\}$ — very tough (only enclosures, often redundant)

Residual values

We will consider minimum-norm estimators:

- $\widehat{\theta}^k := \text{argmin}\|y - X\theta\|_k$ with $k \in \{1, 2, \infty\}$:
  - $k = 1$: Least Absolute Deviations (LAD), can be written as a linear program
  - $k = 2$: Ordinary Least Squares (OLS), can be written explicitly
  - $k = \infty$: Chebyshev Approximation, can be written as a linear program
- The residual value: $R^k = \|y - X\widehat{\theta}^k\|_k$
- **Main goal:** to compute $\underline{R^k}, \overline{R^k}$ for $X \in \mathbf{X}, y \in \mathbf{y}$

Complexity of computation of the residual values

Case:	I	II	III	IV	V
p	unbounded	unbounded	$O(1)$	unbounded	$O(1)$
$\mathbf{X}$	interval	interval	interval	$\underline{X} = \overline{X}$	$\underline{X} = \overline{X}$
$\mathbf{y}$	interval	interval	interval	interval	interval
θ	$\theta \in \mathbb{R}^p$	$\theta \geqslant 0$	$\theta \in \mathbb{R}^p$	$\theta \in \mathbb{R}^p$	$\theta \in \mathbb{R}^p$
$\overline{R}^1$	NPH	NPH	P	P	P
$\underline{R}^1$	NPH	P	P	P	P
$\overline{R}^2$	NPH	NPH	NPH	NPH	NPH
$\underline{R}^2$	NPH	P	P	P	P
$\overline{R}^\infty$	NPH	NPH	P	P	P
$\underline{R}^\infty$	NPH	P	P	P	P

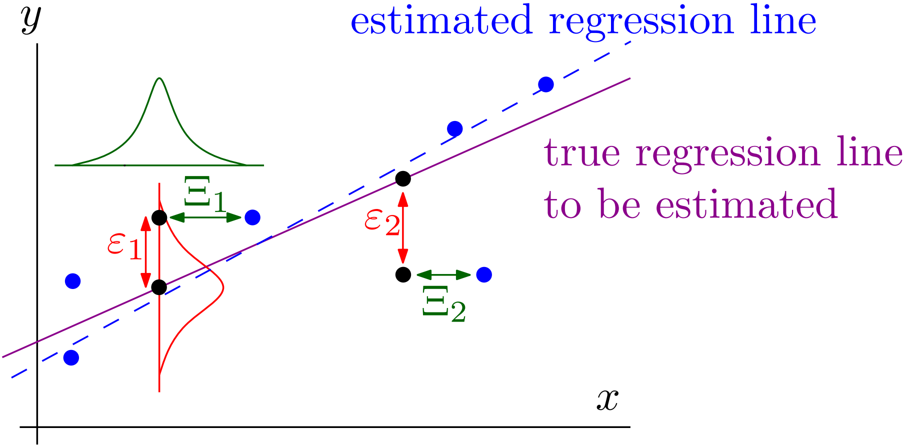
Proof idea: Use the orthant decomposition of the parameter space $\mathbb{R}^p$ and Oettli-Prager Theorem

An application of interval methods in statistics: EIV regression models

EIV regression models

- **Now:** forget intervals — the setup is entirely probabilistic
- Regression model
$$y = X\theta + \varepsilon;$$
observable data are (y, Z), where
$$Z = X + \Xi;$$
here, ε_i's are random errors in (observations of) the dependent variable and Ξ_{ij}'s are random errors in (observations of) regressors. Moreover, X can be taken as a random matrix.
- Since we observe regressors with errors, we speak about Errors-In-Variables (EIV).

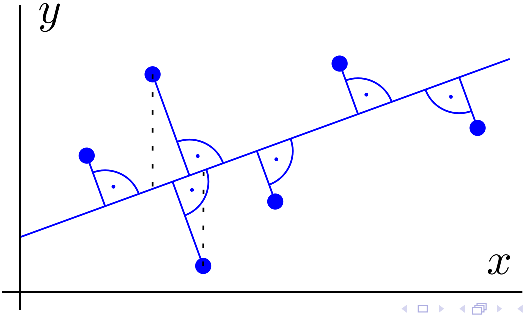
EIV regression models



Total Least Squares

Under traditional assumptions: say, all errors are independent and $N(0,\sigma^2)$ — a “good estimator” is Total Least Squares (TLS):

- Find $\Delta Z, \Delta y, \widehat{\theta}$ s.t.
  - $(Z-\Delta Z)\widehat{\theta} = y - \Delta y$ and
  - $\|(\Delta Z, \Delta y)\|_F$ is minimal, where $\|Q\|_F = \sqrt{\sum_{ij} Q_{ij}^2} = \sqrt{\mathrm{trace}(Q^{\mathrm{T}}Q)}$ is the Frobenius norm
- Then: $\widehat{\theta}$ is a “good” estimate of θ; Δy is an estimate of ε and ΔZ is an estimate of Ξ



Now: Change the matrix norm!

Let's change the assumptions on the error distributions:

- Let all errors have a bounded distribution with support $(-\gamma, +\gamma)$, where $\gamma > 0$ is an unknown constant
- Assume that asymptotically, when $n \to \infty$, the errors approach the bounds $\pm\gamma$ arbitrarily close with $\Pr \to 1$
- Interesting: no independence, zero means or id assumptions are needed
- **Theorem.** Replace the Frobenius norm by Chebyshev norm and you get a consistent estimator.

To compute the estimator, we are to solve the Chebyshev Norm Problem (CNP):

- Find $\Delta Z, \Delta y, \widehat{\theta}$ s.t.
  - $(Z - \Delta Z)\widehat{\theta} = y - \Delta y$ and
  - $\|(\Delta Z, \Delta y)\|_{\max}$ is minimal, where $\|Q\|_{\max} = \max_{ij} |Q_{ij}|$ is the Chebyshev norm

Solving CNP via Oettli-Prager

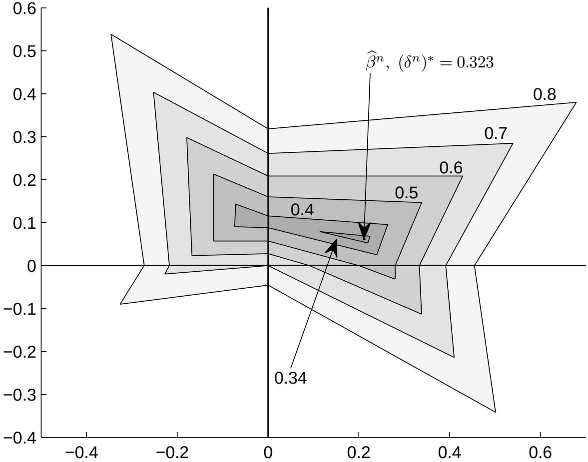
- (CNP) Find $\Delta Z, \Delta y, \widehat{\theta}$ s.t.
  - $(Z - \Delta Z)\widehat{\theta} = y - \Delta y$ and
  - $\|(\Delta Z, \Delta y)\|_{\max}$ is minimal, where $\|Q\|_{\max} = \max_{ij} |Q_{ij}|$ is the Chebyshev norm
- **Equivalently:** Find the minimum δ s.t. the interval-valued linear system
$$[Z \pm \delta E]x = [y \pm \delta e]$$
is solvable (here E is all-one matrix and e is all-one vector).
- Now Oettli-Prager helps: the solution set is a union of polyhedra, convex in each orthant; the polyhedra are parametrized by δ:

$$\begin{aligned} \text{solution set} &= \{x : |Zx - y| \leqslant \delta E|x| + \delta e\} \\ &= \bigcup_{s \in \{\pm 1\}^p} \left\{ x : \begin{array}{rl} Zx - y & \leqslant \delta E D_s x + \delta e, \\ Zx - y & \geqslant -\delta E D_s x - \delta e, \\ D_s x & \geqslant 0 \end{array} \right\}, \end{aligned}$$

where $D_s = \mathrm{diag}(s)$.

Solving CNP via Oettli-Prager: Example

$$Z = \begin{pmatrix} 3 & -0.5 \\ 0.5 & 3 \\ 0.6 & 3 \end{pmatrix}, \qquad y = \begin{pmatrix} 0.2 \\ 0.7 \\ -0.1 \end{pmatrix}.$$



And the last step is easy...

Just rewrite the solution set

$$\bigcup_{s\in\{\pm1\}^p}\left\{x:\ \begin{array}{rl} Zx-y & \leqslant \delta ED_s x+\delta e,\\ Zx-y & \geqslant -\delta ED_s x-\delta e,\\ D_s x & \geqslant 0 \end{array}\right\}$$

as

$$\bigcup_{s\in\{\pm1\}^p}\left\{x:\ \begin{array}{rll} \frac{z_i^{\mathrm{T}}x-y_i}{e^{\mathrm{T}}D_s x+1} & \leqslant \delta, & i=1,\ldots,n,\\ \frac{-z_i^{\mathrm{T}}x+y_i}{e^{\mathrm{T}}D_s x+1} & \leqslant \delta, & i=1,\ldots,n,\\ D_s x & \geqslant 0 & \end{array}\right\}$$

Now, in the orthant s, the minimum δ can be found efficiently via the Generalized Linear-Fractional Program

$$\min_{x\in\mathbb{R}^p}\left\{\max_{\substack{i\in\{1,\ldots,n\}\\ k\in\{0,1\}}}\frac{(-1)^{1-k}z_i^{\mathrm{T}}x+(-1)^k y_i}{e^{\mathrm{T}}D_s x+1}\ \middle|\ D_s x\geqslant 0\right\}.$$

To conclude

Summary:

- The consistent estimator reduces to solving 2^p GLFPs ($p =$ number of regression parameters)
- This is good news: the method is not exponential in the number of observations
- In general, CNP is NP-hard, so nothing better can be achieved
- Both the proof of consistence of the estimator and construction of the "efficient" algorithm for its computation require interval methods (The main tool: Oettli-Prager's decomposition of the space of parameters of the regression model)
- Interesting special case: If we know a priori the signs of regression coefficients (say, $\theta \geqslant 0$), then one GLFP suffices!

And the last slide...

Thank you!