Perturbed Datasets Methods for Hypothesis Testing - Distribution assumption free hypothesis testing -

Sándor Kolumbán
BME Department of Automation and Applied Informatics
HAS MTA-BME Control Engineering Research Group
VUB Brussels Department ELEC

Content

- Birds eye view on hypothesis testing
- Recent history of the distribution free results
- Elementary probability
- Data perturbation methods
  - Symmetric noise distributions
  - Exchangeable noise distributions
- Hypothesis testing with mild assumptions:
  - The noise can be expressed
  - The noise distribution is invariant under transformations from a finite symmetry group

Birds eye view on hypothesis testing

- Measurement data generated as
$$y_k = \theta_0 + n_k$$
- Assume that n_k is $\mathcal{N}(0, \sigma^2)$
- Model under test θ
- The goal is to accept or reject the hypothesis

$$\text{“}\mathrm{H}_0\text{: } \theta = \theta_0\text{”}$$

Birds eye view on hypothesis testing

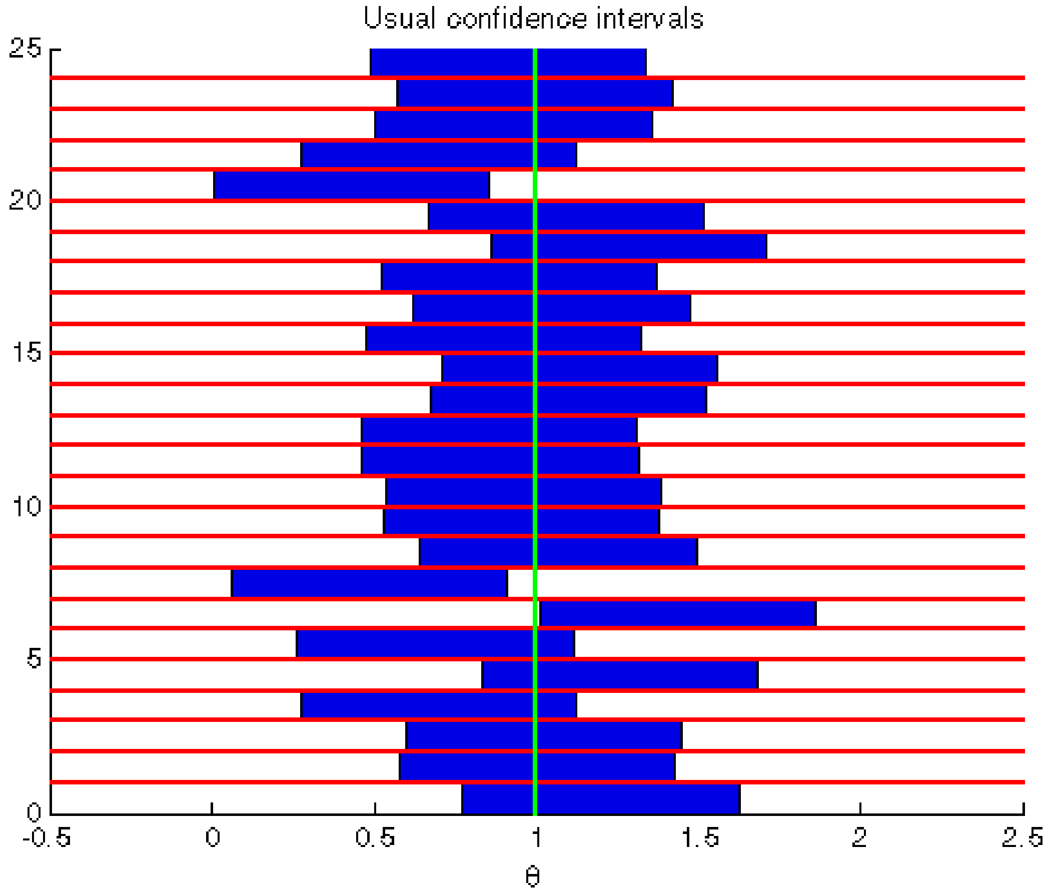
- Create statistic $O(Y, \theta)$, with known distribution
- Select a subset C of the possible outcomes of O where H_0 is accepted
- Let $C_\theta = \{\theta : O(Y, \theta) \in C\}$ be the set of accepted models
- Requirements
  - If$\theta = \theta_0$ then H_0 is accepted with given probability α.
  - If$\theta \neq \theta_0$ then H_0 is rejected with a probability depending on how $\neq$ they are.
  - C_θ should have “nice” properties.

Birds eye view on hypothesis testing

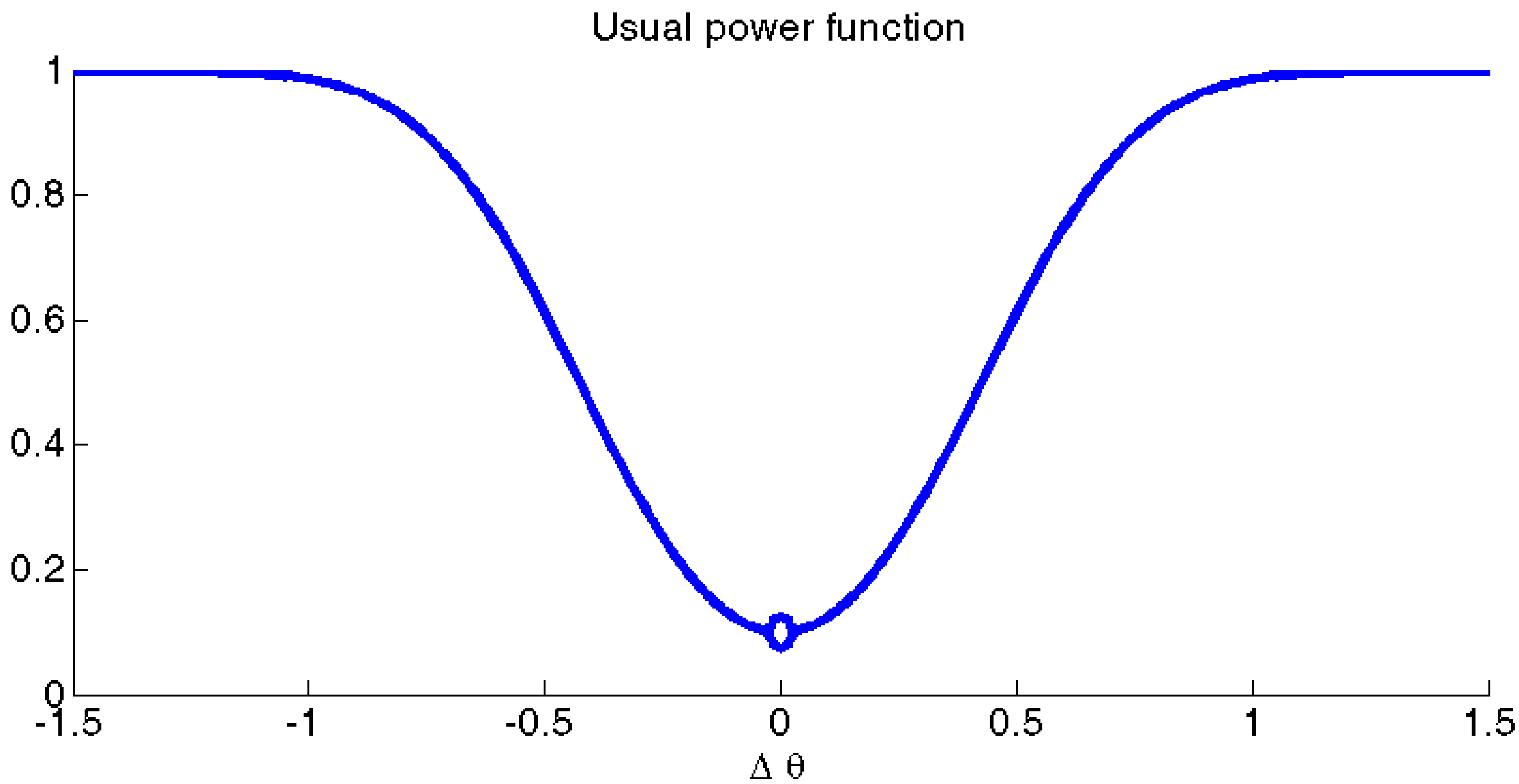
One way to do it (the usual way):

- $O(Y,\theta) = \frac{\sum_{k=1}^{n}(y_k - \theta)}{\sqrt{n\sigma^2}} \sim \mathcal{N}(0,\sigma^2)$
- $C = \left(\Phi^{-1}\left(\frac{1-\alpha}{2}\right), \Phi^{-1}\left(\frac{1+\alpha}{2}\right)\right)$
- $C_\theta = \left(\frac{\sum y_k}{n} - \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(\frac{1+\alpha}{2}\right), \frac{\sum y_k}{n} - \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(\frac{1-\alpha}{2}\right)\right)$

Birds eye view on hypothesis testing



Birds eye view on hypothesis testing

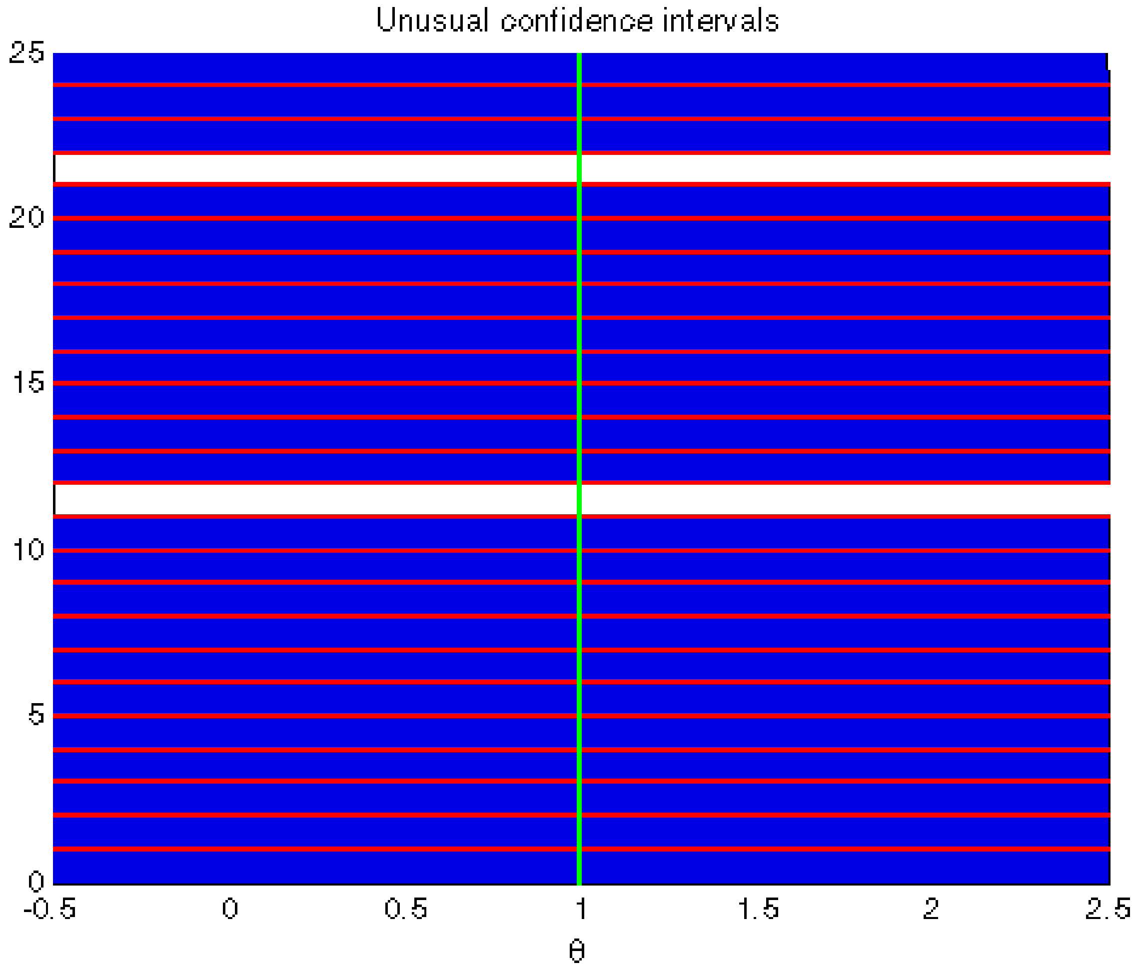


Birds eye view on hypothesis testing

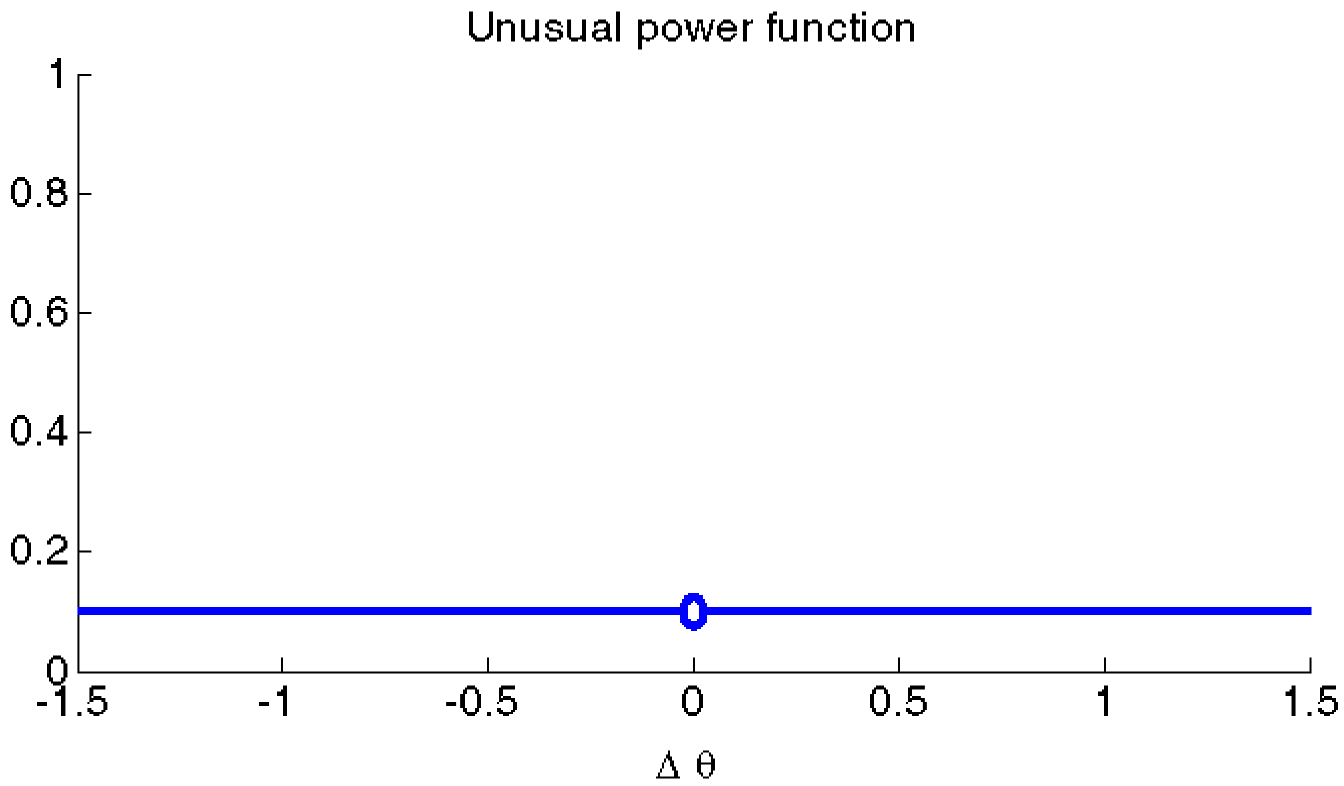
Another way to do it (the unusual way):

- $O(Y,\theta) \sim \Pr(O=1) = \alpha = 1 - \Pr(O=0)$
- $C = \{1\}$
- $C_\theta = 1(O=1)\{-\infty,\infty\} + 1(O=0)\emptyset$

Birds eye view on hypothesis testing



Birds eye view on hypothesis testing



Birds eye view on hypothesis testing

The usual way:

- ✗ Detailed assumptions (model structure, distributions)
- ✓ Deterministic/repeatable decisions (based on observations)
- Central limit theorem + asymptotic theory for the distribution of estimates

The unusual way:

- ✓ No assumptions
- ✗ Totally unrepeatable decisions

Birds eye view on hypothesis testing

The usual way:

- Detailed assumptions (model structure, distributions)
- Deterministic/repeatable decisions (based on observations)
- Central limit theorem + asymptotic theory for the distribution of estimates

It would be nice to meet in the middle!

- Our assumptions are almost never true
- A little bit of coherence in the decisions is desirable
- We aim for exact confidence levels for finite sample count

The unusual way:

- No assumptions
- Totally unrepeatable decisions

Distribution free methods

- The idea was introduced around 2005
- Names: Marco Camp, Balázs Csanád Csáji, Eric Weyer
- Buzz words: LSCR (leave-out sign-dominant correlation regions), SPS (sign-perturbed sums)
- My work:
  - a general framework for distribution free methods (data perturbation methods)
  - SPS is a (meaningful) data perturbation method for linear regression problems with jointly symmetric noise distribution
  - a (meaningful) data perturbation method for linear regression problems with exchangeable noise distribution

Probability basics

- *Randomly well defined ordering*:
  Let π be a uniformly chosen random permutation of $\{1, \ldots, m\}$.
  The well defined ordering by π of a sequence $Z_1, \ldots, Z_m$ is $O_\pi(Z) = [i_1, \ldots, i_m]$ if
  $$Z_{i_1} >_\pi Z_{i_2} >_\pi \cdots >_\pi Z_{i_m}$$
- $\forall\, \pi : Z_i < Z_j \Rightarrow Z_i <_\pi Z_j$
- $Z_i = Z_j \Rightarrow Z_i <_\pi Z_j$ if i precedes j in π

Probability basics

- Almost true: Independent and identically distributed random variables are uniformly ordered.
- If $Z_1, \dots, Z_m$ is an i.i.d. sequence of random variables and π is a uniformly chosen random permutation then $O_\pi(Z)$ is a uniform random permutation

$$\Pr(O_\pi(Z) = [i_1, \dots, i_m]) = \frac{1}{m!}$$

- Proof by symmetry arguments.

Probability basics

- Let $\mathcal{G}(G,\cdot)$ be a finite group, $X_1 = 1, X_{i\geq 2} \sim Uni(G),\ X_0 \sim Uni(G)$, jointly independent.
- If $\tilde{X}_{i\geq 1} = X_i \cdot X_0$ then then $\tilde{X}_{i\geq 1}$are jointly independent and uniformly distributed over G.
- Proof by straight forward calculation.

Probability basics

Groups that will be used

- Sign vectors of length n :

$$G = \{-1,1\}^n$$
$$(x_1 \cdot x_2)[k] = x_1[k]x_2[k]$$

- The symmetric group S_n (group of permutations):

$$x_1 = (3\ 1\ 2), x_2 = (1\ 3\ 2)$$
$$x_1 \cdot x_2 = (2\ 3\ 1)$$

Perturbed dataset methods

- Let the measurements come from a model
$$Y = f(\theta_0, X, N)$$
- f is a known model structure
- X contains the known input values
- N contains disturbing unknown noise
- $\theta_0 \in R^{n_\theta}$ is the parameter vector

Perturbed dataset methods

- Invertibility with respect to noise is required

$$\exists f^*: \Theta \times \mathcal{X} \times \mathcal{Y} \rightarrow \mathcal{N}$$

$$Y = f(\theta, X, N) \Rightarrow N = f^*(\theta, X, Y)$$

- When it is obvious from context

$$N(\theta) = f^*(\theta, X, Y)$$

- The notation D will be used to denote all available data (X and Y usually)

Perturbed dataset methods

- The goal is create a hypothesis test for the parameter vector θ without exact knowledge about the distribution of the noise vector N.
- Some structural symmetry assumptions about the joint distribution of N is required.
- The confidence level can be (almost) arbitrarily selected as $\alpha = \frac{k}{m!}$.
- A random data perturbation setup Γ is required beside the measurements.

Perturbed dataset methods

- Testing θ on confidence level $\alpha = {}^{k}/_{m!}$:
  1. Generate m perturbed datasets $D^{(i)}(D,\theta)$ based on Γ
  2. Define a performance measure $Z\colon\mathcal{D}\times\Theta\to R$
     $$Z_i = Z\left(D^{(i)}(D,\theta),\theta\right)$$
  3. Create a well defined ordering $O_\Gamma(Z)$
  4. Select k out of the possible $m!$ permutations where $H_0\colon\theta=\theta_0$ is considered accepted
- If $\theta = \theta_0$ then Z_i should be i.i.d.

Generating perturbed datasets

- Given X, Y and θ
- Calculate the corresponding noise sequence

$$\hat{N}(\theta) = f^*(\theta, X, Y)$$

- If $\theta = \theta_0$ then $\hat{N}(\theta) = N$
- Create m perturbed noise realization

$$N^{(i)}(\theta, \Gamma) = P_i \hat{N}(\theta)$$

- If $\theta = \theta_0$ then $N^{(i)}(\theta, \Gamma)$ are equally likely nose vectors if the perturbations leaves the noise distribution invariant

Generating perturbed datasets

- Create m perturbed noise realization
$$N^{(i)}(\theta, \Gamma) = P_i \widehat{N}(\theta)$$
- Create m perturbed measurements
$$Y^{(i)} = f\left(\theta, X, N^{(i)}\right)$$
- If $\theta = \theta_0$ then $Y^{(i)}$ are equally likely observations (proof later)
- Γ contains $m-1$ random perturbation objects
- $P_1 = I, Y^{(1)} = Y$

Performance measures

- Given the m equally likely datasets $D^{(i)}(D,\theta)$
- Usual least squares measure

$$\mathrm{Z_i} = J_\theta^{(i)}(\theta) = \frac{1}{n}\sum_{k=1}^{n}\left(f^*\left(\theta, X, Y^{(i)}\right)[k]\right)^2 = \frac{1}{n}||N^{(i)}(\theta)||^2$$

- Sensible measures don’t make sense
  – Noise sequences are equivalent up to measure invariant perturbations
  – Sensible performance measures don’t differentiate between measure invariant points
- Something more sophisticated is needed (see later in concrete case)
- If $\theta = \theta_0$ then the values Z_i are i.i.d.

Creation of the ordering

- Given Z_i and a uniformly chosen random permutation π
- $O = O_\pi(Z)$
- If $\theta = \theta_0$ then the values Z_i are i.i.d. and O is a uniformly distributed random permutation
- Only knowledge about invariant transformations of the noise distribution are needed
- Select k of the $m!$ outcomes of O where $H_0\!: \theta = \theta_0$ is considered accepted

Perturbed dataset – “proof”

- fancy coin toss -

$$\boxed{\{e_1, \dots, e_n\}},\ \boxed{P_0} \rightarrow N = "P_0E" \rightarrow Y = f(\theta_0, X, N)$$

$$\downarrow$$

$$\hat{N}(\theta) = f^*(\theta, X, Y) \qquad (\hat{N}(\theta) = N)$$

$$\boxed{\Gamma = (\pi, [I, P_2, \dots, P_m])}$$

$$I\hat{N} \qquad P_2\hat{N} \qquad \cdots \qquad P_m\hat{N}$$

$$\downarrow \qquad \downarrow \qquad \qquad \downarrow$$

$$"IP_0\hat{E}" \qquad "P_2P_0\hat{E}" \qquad \cdots \qquad "P_mP_0\hat{E}"$$

(i.i.d) wrto

$$\downarrow \qquad \downarrow \qquad \qquad \downarrow$$

$$Y^{(1)} \qquad Y^{(2)} \qquad \cdots \qquad Y^{(m)}$$

$$\downarrow \qquad \downarrow \qquad \qquad \downarrow$$

$$Z_i = Z\left(\theta, X, Y^{(i)}\right) \qquad Z_1 \qquad Z_2 \qquad \cdots \qquad Z_m$$

(i.i.d)

$$\searrow \quad \downarrow \quad \swarrow$$

$$O_\pi(Z)$$

(i.i.d)

Linear regression

$$Y = X^T \theta_0 + N$$

- $Y \in R^n, X \in R^{n_\theta \times n}$ - measured, known
- $N \in R^n$- i.i.d. sequence
  - no symmetry required
  - no moment conditions required
- Goal: create confidence regions for parameter vector θ

Linear regression

- If N is an exchangeable sequence and P is a permutation matrix then $N \approx PN$
- Composition of Γ
  - $P_1 = I, P_{i\geq 2} \sim Uni(S_n)$ – uniform random permutations
  - π uniform random permutation

- Sorry for the abusive notation around permutations and matrices

Linear regression

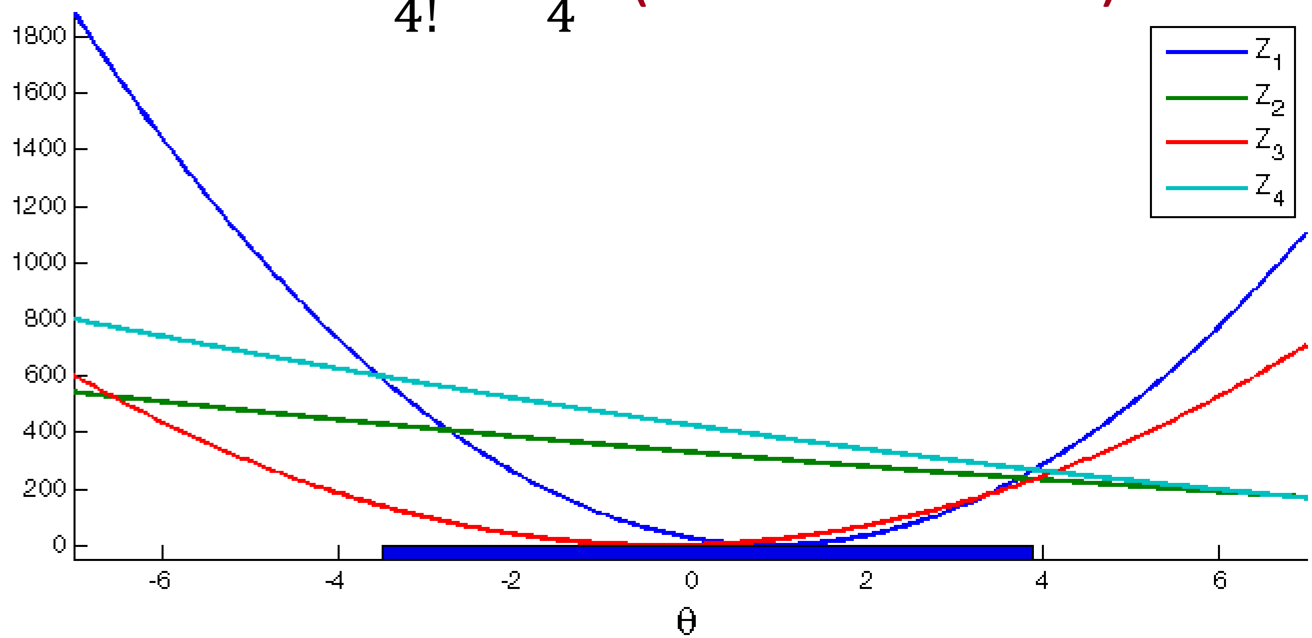
$$Z_i(\theta,\Gamma) = (Y - X^T\theta)^T P_i^T X^T [XX^T]^{-1} X P_i (Y - X^T\theta)$$

- Select orderings such that Z_1 is as small as possible (1 is at the back of the permutation)
- Corresponding confidence regions
  - Contain the least squares estimate
  - Connected
  - Bounded if the input is “exciting enough”
- Proof later, first see a showcase

One dimensional showcase

- $\theta_0 = 1, X = [2,3,-4,-1], N = [4.5,-20,-13,3]^T$
- Γ: $P_2 = (4,1,2,3), P_3 = (2,1,4,3), P_4 = (3,2,1,4)$

$$\alpha = \frac{18}{4!} = \frac{3}{4}$$ (one realization)



Linear regression

- The performance measure is the key
- Perturbed data sets as separate estimation problems
- $J_{\theta}^{(i)}(\theta') = \frac{1}{n}\left(Y^{(i)} - X^T\theta'\right)^T\left(Y^{(i)} - X^T\theta'\right)$
- $\theta^{(i)} = [XX^T]^{-1}XY^{(i)}$ - LS estimate
- $Z_i = \left(\theta^{(i)} - \theta\right)^T[XX^T]\left(\theta^{(i)} - \theta\right)$
- Natural weighting
- $Z_1^{LS} = 0$

Linear regression

- $Z_i =$
$$(Y - X^T\theta_0)^T P_i^T X^T [XX^T]^{-1} X P_i (Y - X^T\theta_0)$$
$$+$$
$$2(Y - X^T\theta_0)^T P_i^T X^T [XX^T]^{-1} X P_i X^T (\theta_0 - \theta)$$
$$+$$
$$(\theta_0 - \theta)^T X P_i^T X^T [XX^T]^{-1} X P_i X^T (\theta_0 - \theta)$$

- $Z_1 - Z_i \approx XX^T - XP_i^T X^T [XX^T]^{-1} X P_i X^T$

Linear regression

- $Z_1 - Z_i \approx XX^T - XP_i^T X^T [XX^T]^{-1} XP_i X^T$
$$=$$
$$X\left(I - P_i^T X^T \left[XP_iP_i^T X^T\right]^{-1} XP_i\right) X^T$$
- I-projection + symmetric sandwich $\Rightarrow$ pos.sem.def.
- Input X is sufficiently exciting with respect to permutation P if
$$XX^T - XP^T X^T [XX^T]^{-1} XPX^T > 0$$

Linear regression

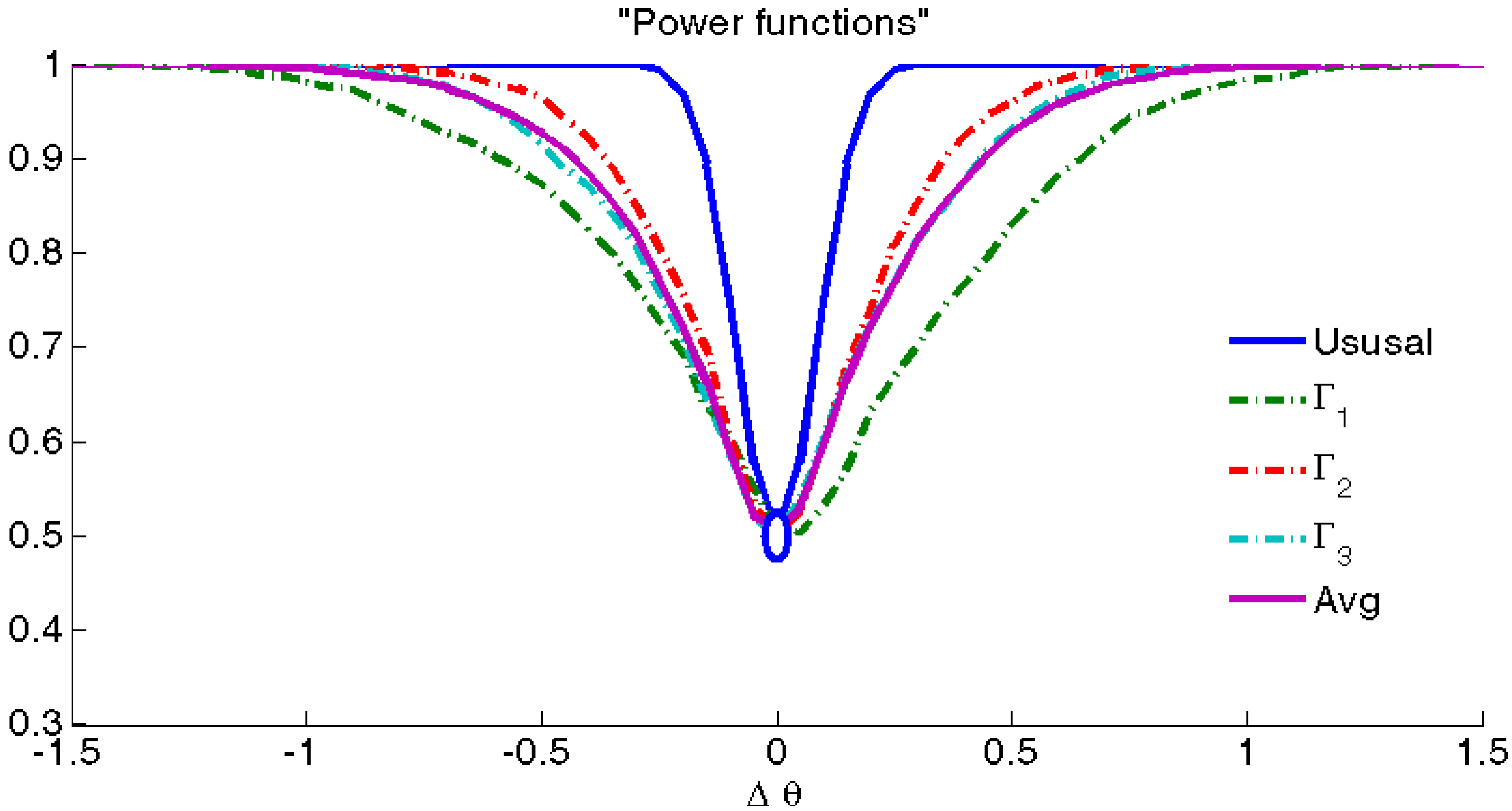
- Input X is sufficiently exciting with respect to permutation P if
$$Q = XX^T - XP^TX^T[XX^T]^{-1}XPX^T > 0$$
- Suff. exc.: $[|\theta_0 - \theta\,| \to \infty] \Rightarrow Z_1 - Z_0 \to \infty$ (power function tends to 1)
- One dimensional constant input – not good enough – $Q = 0$
- Complex problem is required for nontrivial results

Sign-perturbed sums

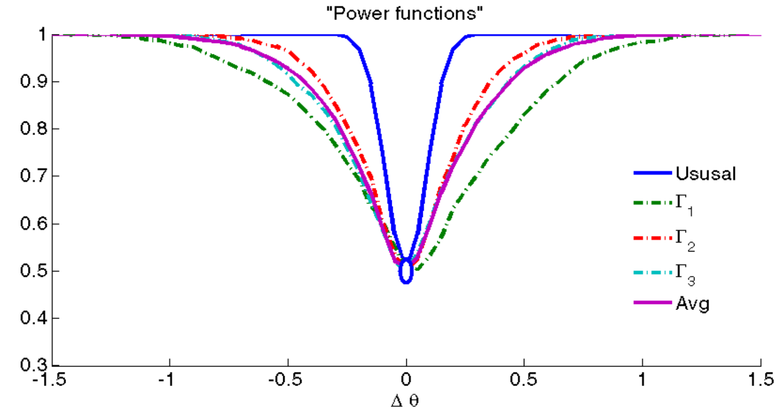
- The method of sign-perturbed sums is also a data perturbation method.
- SPS works with jointly symmetric noise distributions.
- The matrices P_i are not random permutation matrices but diagonal matrices with uniformly distributed random signs $\{-1,1\}$.
- Similar properties for linear regression problems as presented for the i.i.d. case.
- There are two different performance measures Z resulting in nice confidence regions.

Price of information

$$y_k = x_k\theta_0 + e_k,\, e_k \sim \mathcal{N}(0, \sigma^2)$$



Price of information

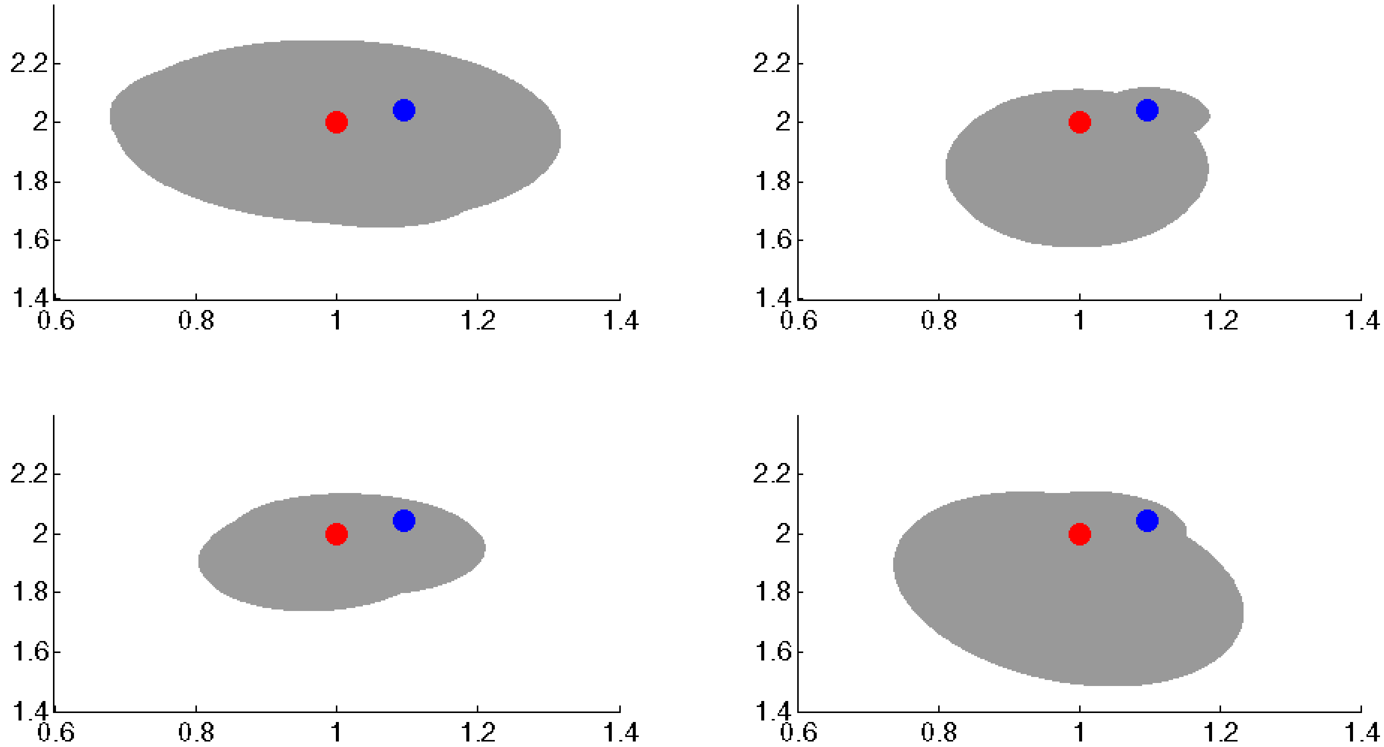


- Minimum at zero because of symmetry
- Loss of power is significant, but the distribution of e_k is not used
- Accurate definition of power function is an issue

Coherence of decisions

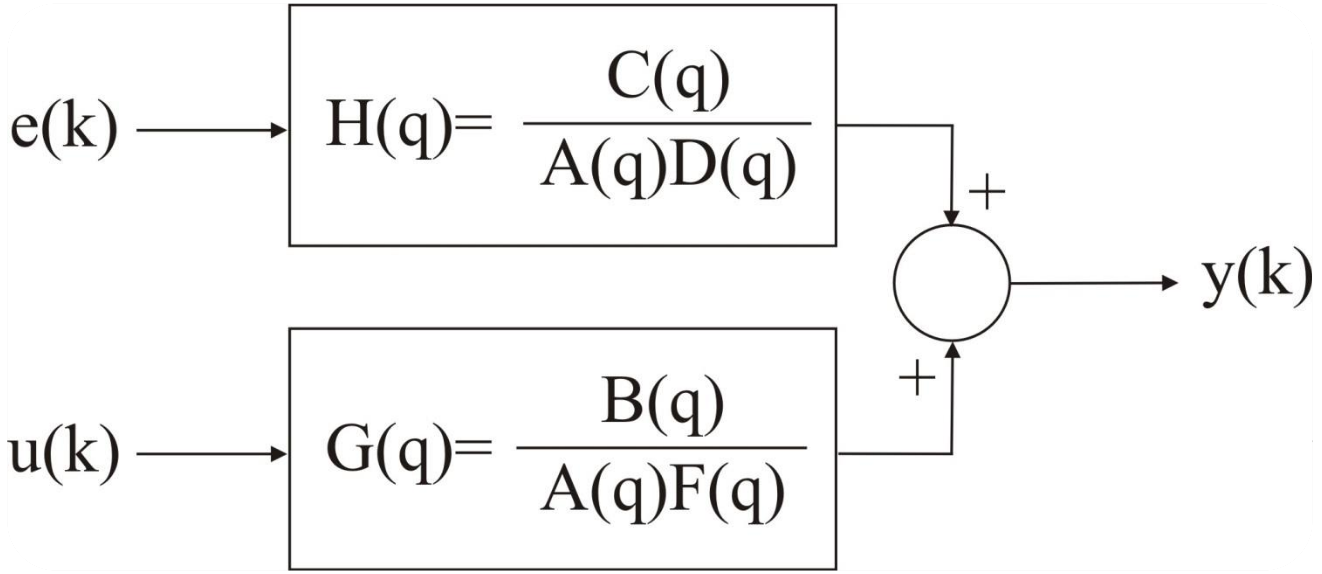
$$y_k = x_k\theta_0 + e_k, e_k \sim Exp(5), n = 25, \alpha = 0.75$$

● θ_0 ● θ^{LS}



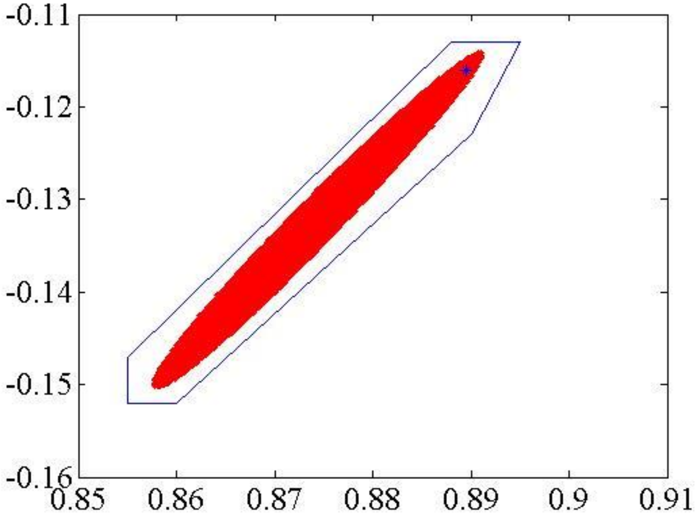
Non linear problems

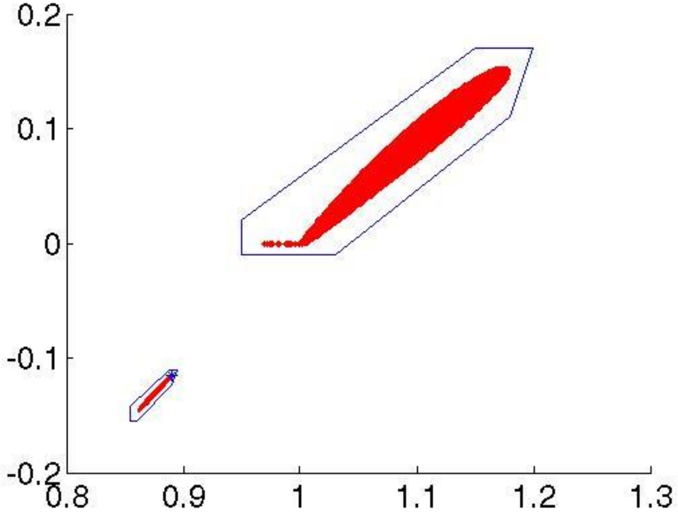
- Confidence regions for parameters of linear dynamical systems



Non linear problems

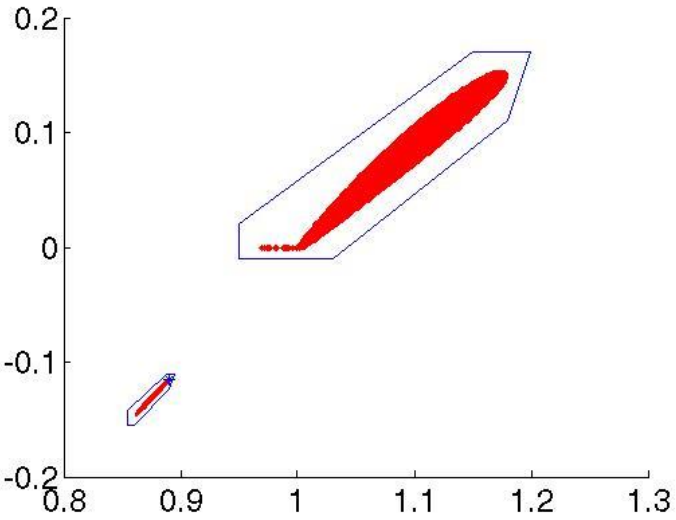
$$\frac{D(q)}{C(q)}\left(A(q)y[k]-\frac{B(q)}{F(q)}u[k]\right)=e[k]$$





Non linear problems

- Uncertainty evaluation is not trivial
- Structural properties depend on problem and performance measure



- Discovering the entire confidence region is hard

Open questions

- The notion of power function is not defined
- Limiting results are not yet proven
- …

Summary

- Hypothesis testing with mild assumptions:
  - The noise can be expressed
  - The noise distribution is invariant under transformations from a finite symmetry group
- The result is random even for fixed observations but not “too random”
- Nice structural results for linear regression problems

Summary

- Hypothesis testing with mild assumptions:
  - The noise can be expressed
  - The noise distribution is invariant under transformations from a finite symmetry group
- The result is random even for fixed observations but not “too random”
- Nice structural results for linear regression problems
- Thank you for your attention!