XIII Colóquio Brasileiro de Ciências Geodésicas • 2024
Universidade Federal do Paraná
25 Anos
Conectando mentes e provendo conhecimento

MENOR ERRO INFLUENCIÁVEL: UMA NOVA MÉTRICA PARA PROTEGER OS PARÂMETROS DE OUTLIERS

Ivandro Klein[1,2], Marcelo Tomio Matsuoka[3], Vinicius Francisco Rofatto[3]

[1] Instituto Federal de Santa Catarina (IFSC)
[2] Universidade Federal do Paraná (UFPR)
[3] Universidade Federal de Uberlândia (UFU)

Curitiba, 26 a 29 de novembro de 2024

Introdução

- **Definição de *outlier***: *"uma observação que se afastou do valor mais provável a ponto de não pertencer ao modelo matemático (funcional e estocástico) estipulado"* (ROFATTO, 2020; ROFATTO *et al.*, 2022)

- ***Outliers* não detectados:** falhas na estimação do modelo que podem resultar até mesmo em **risco a segurança de vidas** e ao **meio ambiente** (exs: monitoramento geodésico de estruturas, navegação com GNSS etc.)

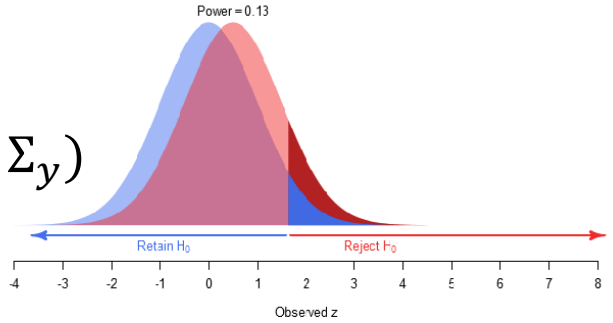
- **Identificação de *outliers***: em geral analisando o **tamanho do resíduo normalizado** das observações (Baarda's *Data Snooping – DS*, Pope's *Tau Test*, estimadores robustos reponderados etc.)

- **Nossa proposta**: analisar a **distribuição multivariada dos resíduos** bem como as **regiões de confiança** desejadas para os **parâmetros**, por meio do novo conceito de "**menor erro influenciável**" das observações

O modelo nulo (m_0) e seus candidatos alternativos (m_i)

- **Modelo nulo** (dados com erros aleatórios normalmente distribuídos) $m_0: \underline{l} \sim \mathcal{N}(A\delta x, \Sigma_y)$

- **Modelo alternativo** (*outlier* de valor ∇_i na i-ésima observação) $m_i: \underline{l} \sim \mathcal{N}(A\delta x + c_i \nabla_i, \Sigma_y)$

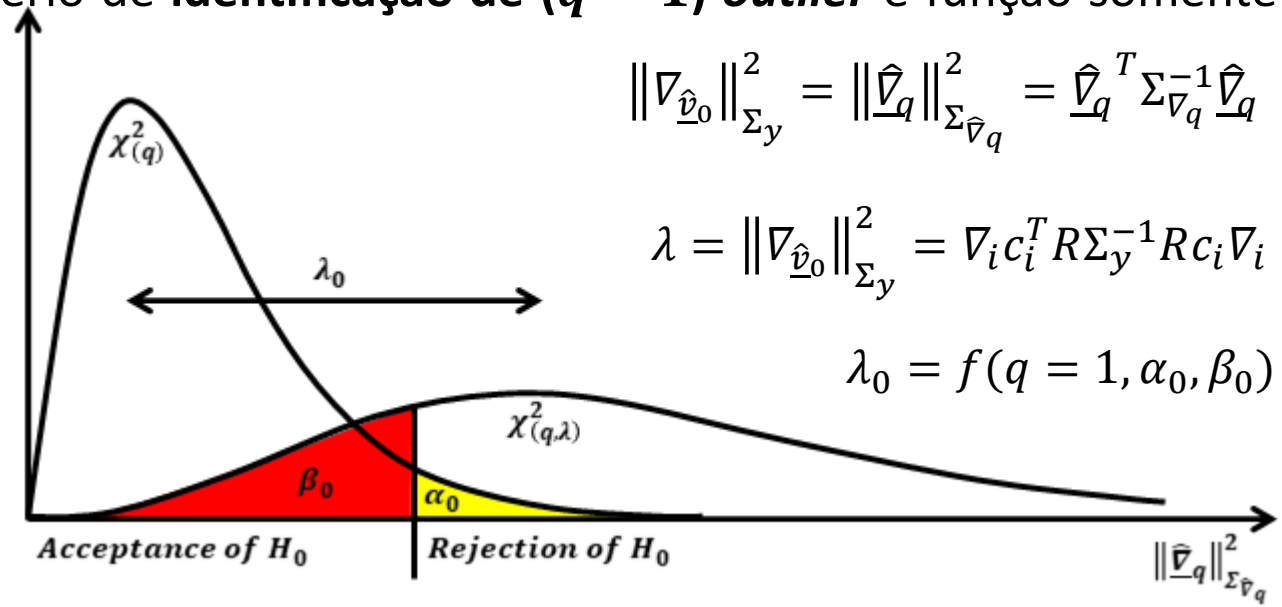


- **Abordagem convencional**: ∇_i é **desconhecido** e o critério de **identificação de ($q = 1$) *outlier*** é função somente do **tamanho dos resíduos padronizados**

$$\left\|\nabla_{\underline{\hat{v}}_0}\right\|^2_{\Sigma_y} = \left\|\underline{\hat{\nabla}}_q\right\|^2_{\Sigma_{\hat{\nabla}_q}} = \underline{\hat{\nabla}}_q{}^T \Sigma^{-1}_{\nabla_q} \underline{\hat{\nabla}}_q$$

$$\lambda = \left\|\nabla_{\underline{\hat{v}}_0}\right\|^2_{\Sigma_y} = \nabla_i c_i^T R \Sigma_y^{-1} R c_i \nabla_i$$

$$\lambda_0 = f(q = 1, \alpha_0, \beta_0)$$



- *Baarda's* ***Minimal Detectable Bias*** (MDB):

$$\nabla_i{}^2 = \lambda_0 \left(c_i^T R \Sigma_y^{-1} R c_i\right)^{-1} \Rightarrow |\nabla_i| = \sqrt{\frac{\lambda_0}{c_i^T R \Sigma_y^{-1} R c_i}}$$

Novo conceito: o menor erro influenciável (MEI) de cada observação

- **MEI**: tamanho do *outlier* que induz um **bias significativo** nos **parâmetros** estimados (quando o **tamanho** e a **direção** de sua **projeção** no espaço dos parâmetros **extrapolam a região de confiança** desejada)

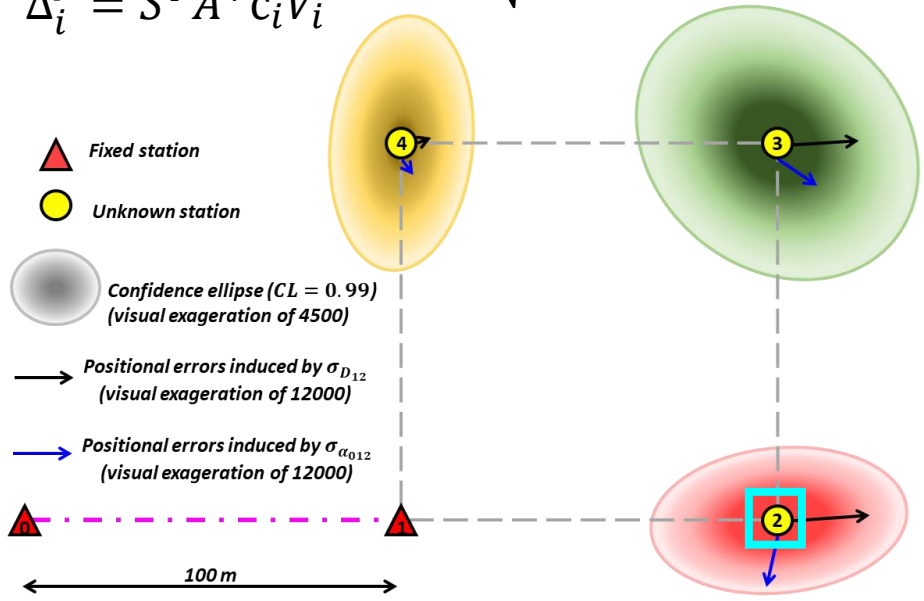
$$\underline{\hat{\Delta}}_x = \left(A^T\Sigma_y^{-1}A\right)^{-1}A^T\Sigma_y^{-1}\underline{\epsilon} = A^+\underline{\epsilon} \qquad \left\|\Delta_i^p\right\|^2_{S^T\Sigma_{\hat{x}}S} = \nabla_i^p c_i^T A^{+T} S(S^T\Sigma_{\hat{x}}S)^{-1}S^TA^+c_i\nabla_i^p = \chi^2_{(d,\alpha)} \qquad |\nabla_i| = min\left|\nabla_i^p\right| (p = 1, \dots, m)$$

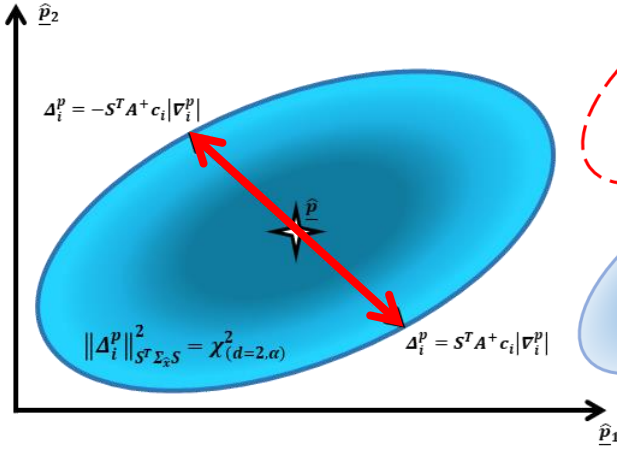
$$\epsilon = c_i\nabla_i \qquad \left|\nabla_i^p\right| = \sqrt{\frac{\chi^2_{(d,\alpha)}}{c_i^T A^{+T} S(S^T\Sigma_{\hat{x}}S)^{-1}S^TA^+c_i}} \qquad \Delta_i^p = S^TA^+c_i\nabla_i$$

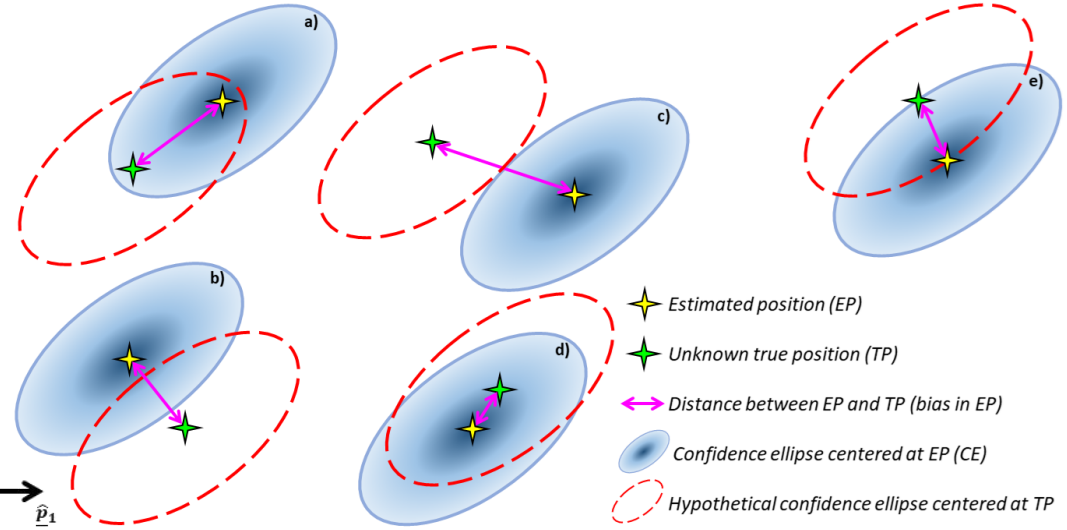
Ao invés de fixar $\boldsymbol{\nabla_i = MDB}$ e obter $\underline{\hat{\Delta}}_x$,

obtém-se $\boldsymbol{\nabla_i = MEI}$ condicionando $\underline{\hat{\Delta}}_x$ $\quad S^T = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$

Teunissen (2018): como otimizar a regra de decisão via $\underline{\hat{\Delta}}_x$?!







Regra de decisão proposta para seleção do modelo: médias deslocadas

$$d_{(\underline{\hat{v}}_0,m_i)} = \left\|\underline{\hat{v}}_0 - \mu_i\right\|^2_{\Sigma_y} \quad Select\ m_i\ if\ \left\|\underline{\hat{v}}_0 - \mu_i\right\|^2_{\Sigma_y} = min,\ \ i = 0,1,\dots,n$$

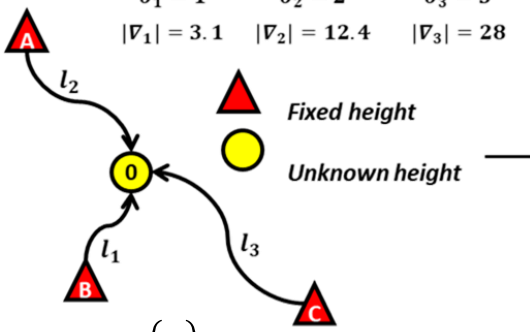
$$d_{(\underline{\hat{v}}_0,m_i)} = \left(\underline{\hat{v}}_0 - \mu_i\right)^T \Sigma_y^{-1}\left(\underline{\hat{v}}_0 - \mu_i\right)$$

a)

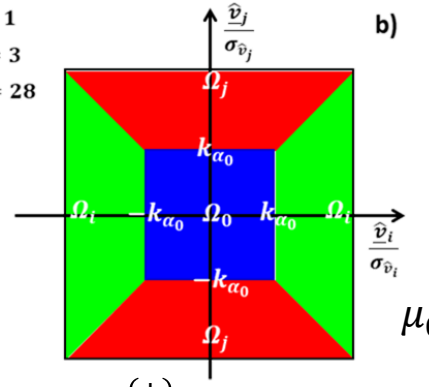
$CL = 0.95$ $\quad n = 3$ $\quad u = 1$

$\sigma_1 = 1$ $\quad \sigma_2 = 2$ $\quad \sigma_3 = 3$

$|\nabla_1| = 3.1$ $\quad |\nabla_2| = 12.4$ $\quad |\nabla_3| = 28$



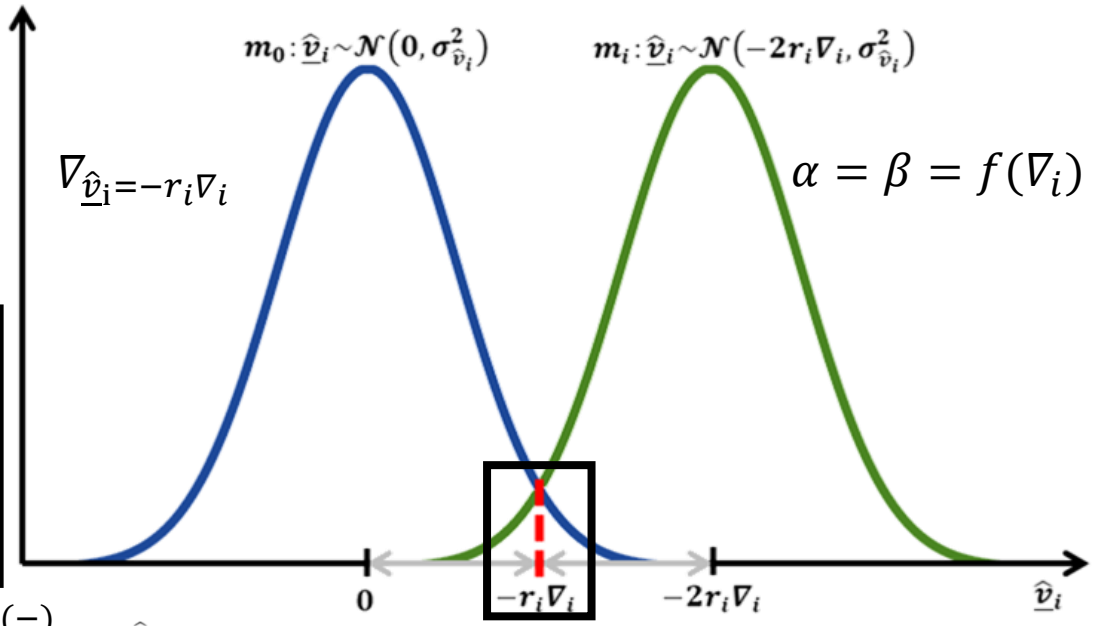
$m_i = m_2^{(-)}$



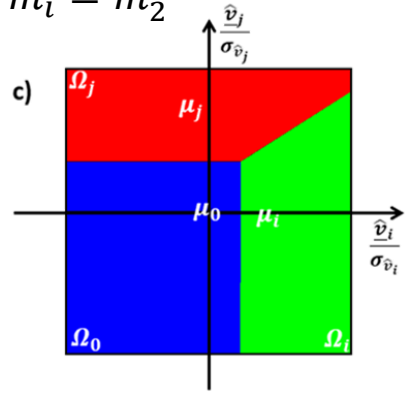
$m_j = m_3^{(+)}$

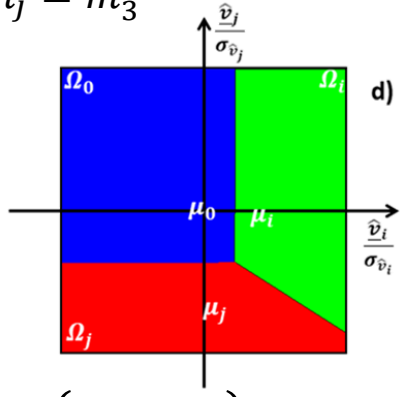
$$\mu_i^{(+)} = \begin{bmatrix} -2r_{1,i}|\nabla_i| \\ \vdots \\ \underbrace{-2r_{i,i}|\nabla_i|}_{i} \\ \vdots \\ -2r_{n,i}|\nabla_i| \end{bmatrix} \qquad \mu_i^{(-)} = \begin{bmatrix} 2r_{1,i}|\nabla_i| \\ \vdots \\ \underbrace{2r_{i,i}|\nabla_i|}_{i} \\ \vdots \\ 2r_{n,i}|\nabla_i| \end{bmatrix}$$

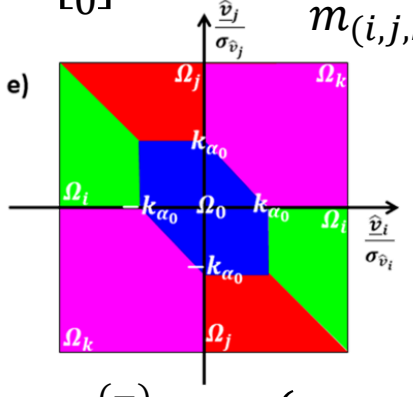
$$\mu_0 = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}$$

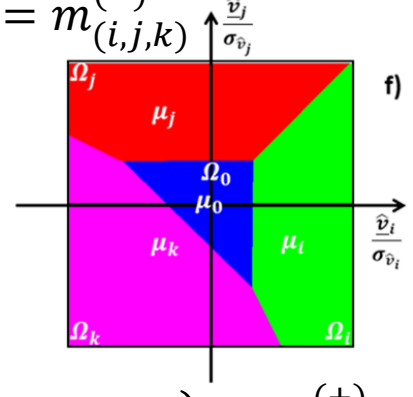


$$m_{(i,j,k)} = m^{(-)}_{(i,j,k)}$$









$$\sigma_i = \sigma_j = \sigma_k = \sigma_0$$

$$\underline{\hat{v}}_i + \underline{\hat{v}}_j + \underline{\hat{v}}_k = 0$$

$$\underline{\hat{v}}_k = -\left(\underline{\hat{v}}_i + \underline{\hat{v}}_j\right)$$

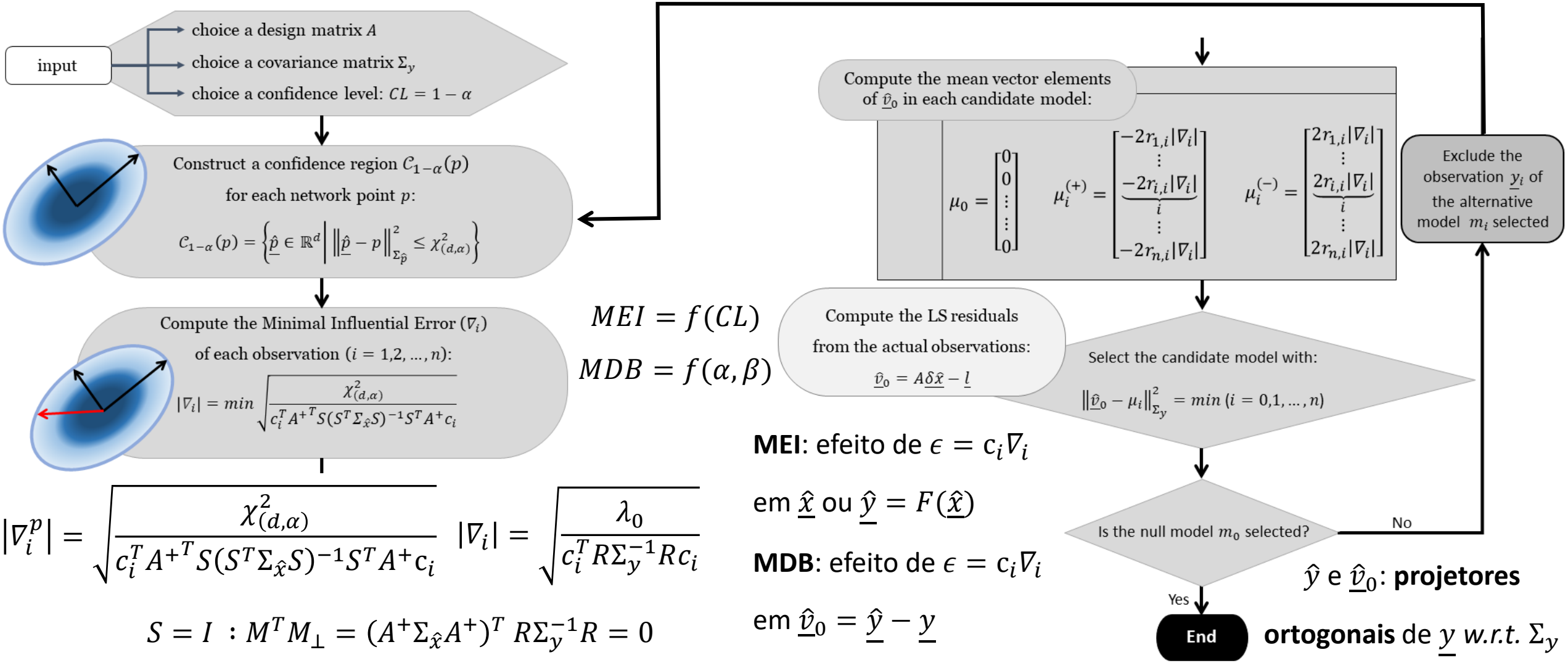
$$\left|\underline{\hat{v}}_k\right| = \left|\underline{\hat{v}}_i + \underline{\hat{v}}_j\right|$$

Análise multivariada do vetor $\underline{\hat{v}}_0$!

$m_j = m_3^{(-)}$

$$m_0: \underline{l} \sim \mathcal{N}\left(A\delta x, \Sigma_y\right) for\ i = 0; \quad m_i^{(-)}: \underline{l} \sim \mathcal{N}\left(A\delta x - 2c_i|\nabla_i|, \Sigma_y\right) or\ m_i^{(+)}: \underline{l} \sim \mathcal{N}\left(A\delta x + 2c_i|\nabla_i|, \Sigma_y\right) for\ 1 \le i \le n$$

Fluxograma da nova abordagem proposta para identificação de *outliers*



$$|\nabla_i^p| = \sqrt{\frac{\chi^2_{(d,\alpha)}}{c_i^T A^{+T} S (S^T \Sigma_{\hat{x}} S)^{-1} S^T A^+ c_i}} \qquad |\nabla_i| = \sqrt{\frac{\lambda_0}{c_i^T R \Sigma_y^{-1} R c_i}}$$

$$S = I \ : M^T M_\perp = (A^+ \Sigma_{\hat{x}} A^+)^T \, R \Sigma_y^{-1} R = 0$$

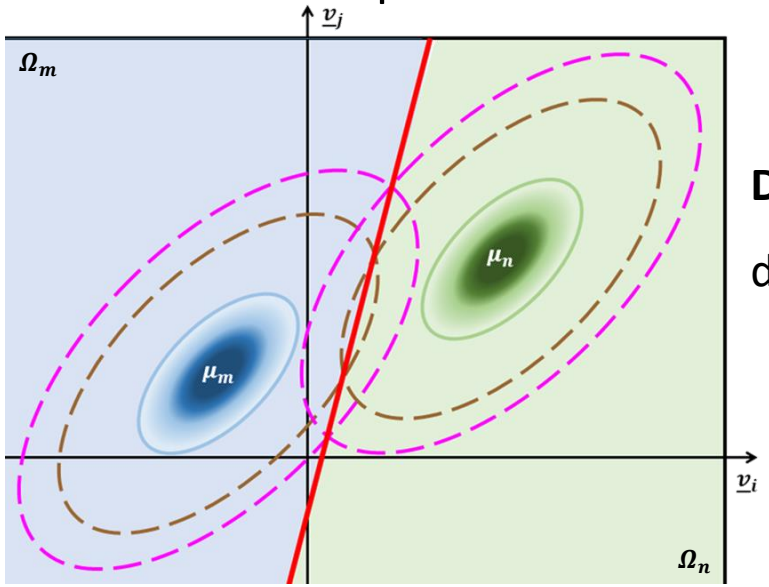
MEI: efeito de $\epsilon = c_i \nabla_i$ em $\hat{\underline{x}}$ ou $\hat{\underline{y}} = F(\hat{\underline{x}})$

MDB: efeito de $\epsilon = c_i \nabla_i$ em $\hat{\underline{v}}_0 = \hat{\underline{y}} - \underline{y}$

$\hat{y}$ e $\hat{\underline{v}}_0$: **projetores ortogonais** de $\underline{y}$ *w.r.t.* Σ_y

Algumas considerações sobre os níveis de probabilidades

- **Níveis de probabilidade** como falsos positivos (**α**) e falsos negativos (**β**) **não** são **arbitrados** (conhecidos *a priori*), mas **limiares** para estes **erros de decisão** podem ser obtidos pela **Distância de Bhattacharyya** (BD)



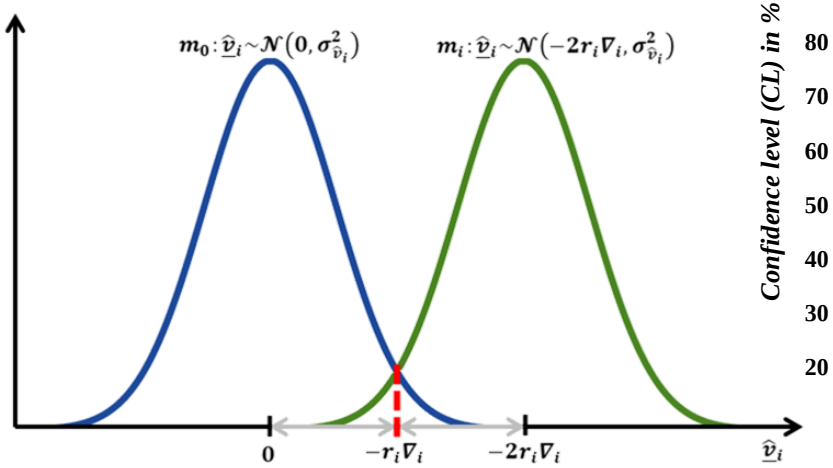
$$BD_{(m,n)} = \frac{1}{8}(\mu_m - \mu_n)^T \left(\frac{\Sigma_m + \Sigma_n}{2}\right)^{-1} (\mu_m - \mu_n)$$

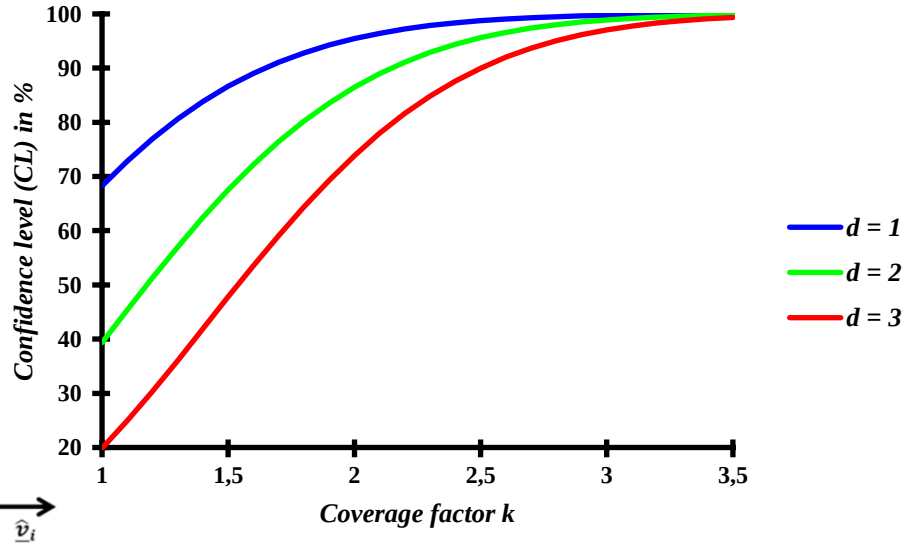
Diferença em μ_m e μ_n ($\Sigma_m = \Sigma_n$): erros de decisão iguais entre dois modelos m_m e m_n !

$$\iint_{\Omega_n} p(\underline{\hat{v}}_0 | m_m) d\underline{\hat{v}}_0 \leq \frac{1}{2} e^{-BD_{(m,n)}}$$

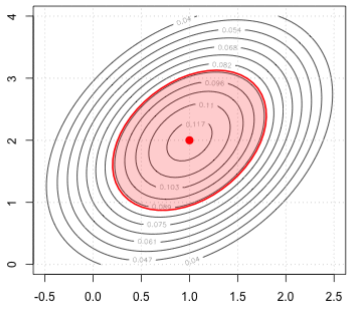
Nível de confiança (CL) *vs.* expansão da região de confiança padrão (fator k) e dimensão espacial da rede (d)

$$CL = \frac{\int_0^{k^2} e^{\left(\frac{-t}{2}\right)} dt}{2\int_0^{\infty} e^{-t} dt} = \frac{1}{2}\left(2 - 2e^{\frac{-k^2}{2}}\right) = 1 - e^{\frac{-k^2}{2}}$$



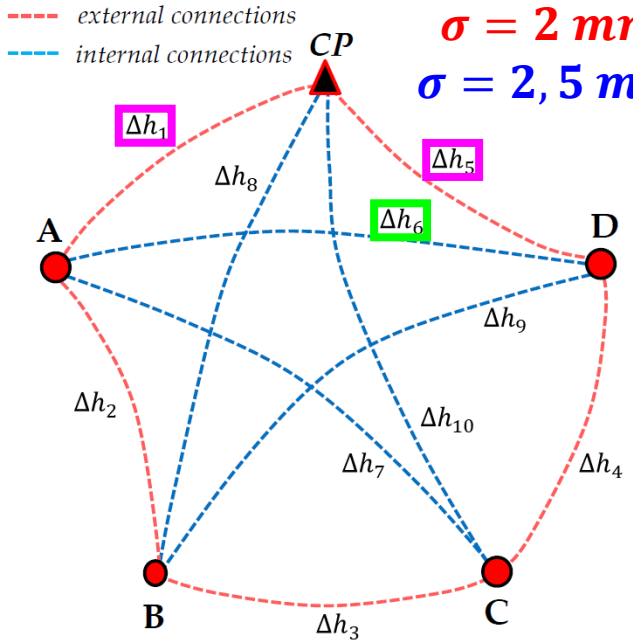


$$\alpha = \beta = f(\nabla_i)$$



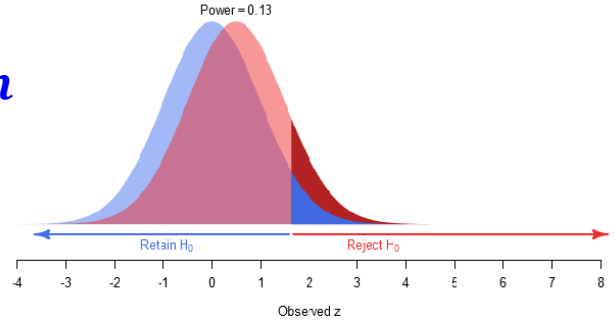
Monte-Carlo (MC): probabilidades de "falsos +" e "falsos –"

Experimentos numéricos: rede de nivelamento geométrico



$\sigma = 2\ mm$

$\sigma = 2,5\ mm$



Probabilidades não simétricas:

Múltiplos modelos ao invés do caso binário m_n e m_m!

$n = 10, u = 4, d = 1$

- **MDB**: 11,2 mm ($\sigma = 2$ mm) ou 12,7 mm ($\sigma = 2{,}5$ mm)
- **MEI**: 5,5 mm (Δh_1 ou Δh_5) até 16,7 mm (Δh_6)
- **Menor MEI** = **Maior** probabilidade de **falso negativo**
- **Exemplo de otimização**: $\sigma = 3$ mm para h_1 e Δh_5
- **Erros de decisão menores** que os limiares por **BD**!

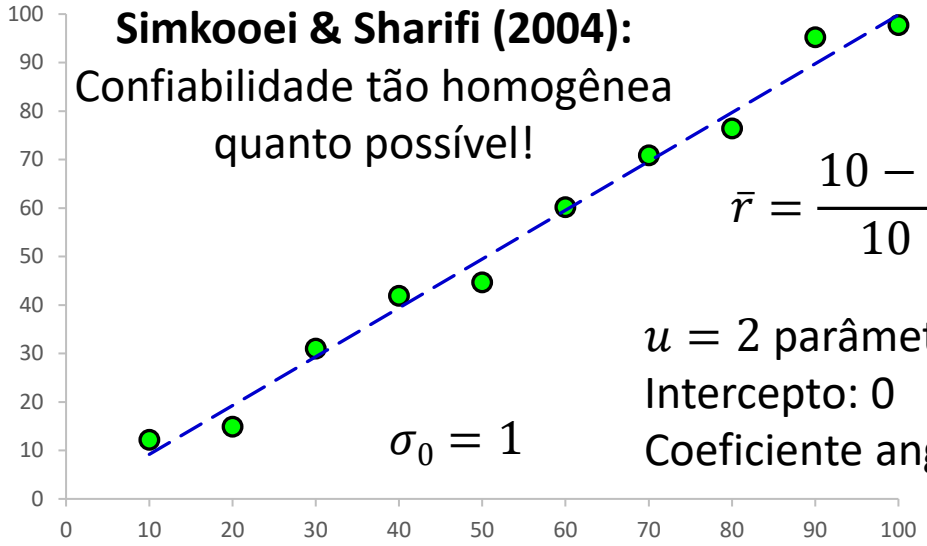
Table 1 – Values of σ, MDB ($\alpha = 0.1\%, \beta = 20\%$), MIE ($CL = 95\%$) and $\frac{1}{2}\, e^{-BD_{(0,i)}}$ for each observation.

Observation	σ (mm)	MDB (mm)	MIE (mm)	$\frac{1}{2}\, e^{-BD_{(0,i)}}$
Δh_1	2	11.2	5.5	6.29%
Δh_2	2	11.2	10.5	0.03%
Δh_3	2	11.2	11.6	0.01%
Δh_4	2	11.2	10.5	0.03%
Δh_5	2	11.2	5.5	6.29%
Δh_6	2.5	12.7	16.7	0.00%
Δh_7	2.5	12.7	16.0	0.00%
Δh_8	2.5	12.7	8.8	0.83%
Δh_9	2.5	12.7	16.0	0.00%
Δh_{10}	2.5	12.7	8.8	0.83%

Table 2 – Classification results for each candidate model by (200,000) MC simulations.

Selected model (in %) \ *"Unknown reality"*	m_0	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9	m_{10}
m_0	**91.593**	1.939	0.003	0.002	0.002	1.856	0.000	0.000	0.156	0.000	0.155
m_1	3.815	**96.532**	0.018	0.000	0.001	1.233	0.000	0.000	0.173	0.000	0.132
m_2	0.009	0.018	**99.975**	0.000	0.000	0.000	0.000	0.000	0.004	0.000	0.000
m_3	0.002	0.000	0.001	**99.995**	0.000	0.000	0.000	0.000	0.000	0.000	0.001
m_4	0.007	0.000	0.000	0.001	**99.982**	0.017	0.000	0.000	0.000	0.000	0.006
m_5	3.851	1.191	0.001	0.000	0.014	**96.575**	0.000	0.000	0.130	0.000	0.168
m_6	0.000	0.000	0.000	0.000	0.000	0.000	**100.000**	0.000	0.000	0.000	0.000
m_7	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**100.000**	0.000	0.000	0.000
m_8	0.347	0.182	0.003	0.003	0.000	0.136	0.000	0.000	**99.521**	0.000	0.019
m_9	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	**100.000**	0.000
m_{10}	0.377	0.139	0.001	0.001	0.003	0.184	0.000	0.000	0.017	0.000	**99.521**

Experimentos numéricos: regressão linear com $n = 10$ pontos



Simkooei & Sharifi (2004):
Confiabilidade tão homogênea quanto possível!

$$\bar{r} = \frac{10-2}{10} = 0{,}8$$

$\sigma_0 = 1$

$u = 2$ parâmetros:
Intercepto: 0
Coeficiente angular: 1

Table 3 – Success rate (in %) of each outlier identification scenario by (200,000) MC simulations.

Observation (linear regression point)	Outlier interval $(3-6)\sigma_0$		Outlier interval $(6-9)\sigma_0$	
	Conventional DS	Proposed approach	Conventional DS	Proposed approach
1	*62.8635*	*70.2420*	*99.0110*	*99.4435*
2	*69.9960*	*70.9765*	*99.5970*	*99.6210*
3	*74.3055*	*71.5550*	*99.8305*	*99.7655*
4	*77.2775*	*72.3575*	*99.9015*	*99.8155*
5	*78.5035*	*72.4900*	*99.9210*	*99.8265*
6	*78.3230*	*72.3440*	*99.9155*	*99.8335*
7	*77.1435*	*72.0250*	*99.8105*	*99.8085*
8	*74.5050*	*71.7425*	*99.8270*	*99.7510*
9	*69.9395*	*70.9740*	*99.6170*	*99.6270*
10	*62.8405*	*70.3770*	*99.0170*	*99.4705*
Mean:	***72.5698***	***71.5084***	***99.6448***	***99.6963***

- $d = u = 2$ $(S = I)$ e $CL = 0.9$ (90%): $MEI = 3{,}65\ \sigma_0$
- α **por MC**: $1{,}33\%$ (valor critico do DS por MC: $k = 3{,}21$)
- ***Outliers* moderados**: diferenças significativas para o DS
- ***Outliers* elevados**: diferenças negligenciáveis para o DS
- Nossa abordagem tem **identificação mais homogênea**
- Nossa abordagem **identifica melhor** quando $r_i < \bar{r}$

Table 4 – Redundancy number and PFA (in %) for each observation in both approaches.

Observation (linear regression point)	Redundancy number (r_i)	Observation identified (no outlier case)	
		Conventional DS	Proposed approach
1	*0.655*	*0.126*	*0.3125*
2	*0.752*	*0.127*	*0.145*
3	*0.824*	*0.127*	*0.084*
4	*0.873*	*0.1425*	*0.069*
5	*0.897*	*0.1305*	*0.0505*
6	*0.897*	*0.1355*	*0.0555*
7	*0.873*	*0.134*	*0.0605*
8	*0.824*	*0.1165*	*0.082*
9	*0.752*	*0.131*	*0.1555*
10	*0.655*	*0.122*	*0.2945*
	Total:	***1.292***	***1.309***

Considerações finais

- Nossa abordagem **não requer a probabilidade de falsos positivos** (α) do teste e nem de **falsos negativos** (β) para as medidas de confiabilidade, apenas o **nível de confiança** (CL) das **regiões de confiança** dos parâmetros

- O novo conceito proposto de **"Menor Erro Influenciável"** (MEI) **difere** do conceito clássico de **"Menor Erro Detectável"** (MDB), pois considera os efeitos dos *outliers* em $\hat{x}$ ou $\hat{y} = f(\hat{x})$ ao invés de analisar somente $\hat{v}_0$

- Em outras palavras: nossa proposta busca pelo "***outlier* mais influente**" (considerando o impacto nos parâmetros $\hat{x}$) **ao invés do "*outlier* mais provável"** (impacto apenas nos resíduos $\hat{v}_0$: maior resíduo normalizado como DS)

- **Yang et al. (2021)**: atualmente o maior desafio ocorre para ***outliers* de magnitude "moderada"** em **"geometria pobre"** (pelo conceito do MEI, nossa abordagem identifica mais *outliers* "moderados" quando $r_i < \bar{r}$)

XIII Colóquio Brasileiro de Ciências Geodésicas • 2024

Universidade Federal do Paraná

25 Anos

Conectando mentes e provendo conhecimento

AGRADECIMENTOS

PPGAIG
Programa de Pós-Graduação em Agricultura e Informações Geoespaciais

AGRIMENSURA
INSTITUTO FEDERAL DE SANTA CATARINA

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

PÓS-GRADUAÇÃO EM CIÊNCIAS GEODÉSICAS

ivandroklein@gmail.com

REALIZAÇÃO