Hidden in plain sight

Influential sets in linear regression

Nikolas Kuschnig*, Gregor Zens, Jesús Crespo Cuaresma

EEA-ESEM Milano, 22nd of August, 2022

Vienna University of Economics and Business
*nikolas.kuschnig@wu.ac.at

The motivation

How sensitive are our inferences to our data?

The motivation

__H__ow sensitive are our inferences to __o__ur data?

What if our results depend on **one** observation?

The motivation

Ow sensitive are our inferences tour data?

What if our results depend on one observation?

- The issue has been studied in detail.

The motivation

How sensitive are our ***infere****nces to our data?*

What if our results depend on **a few** observations?

- For single observations, the issue has been studied in detail.

The motivation

How sensitive are our fences to our data?

What if our results depend on a few observations?

- For single observations, the issue has been studied in detail.
- The issue is not well understood, and quickly intractable.

The motivation

How sensitive are our fences to our data?

What if our results depend on a few observations?

- For single observations, the issue has been studied in detail.
- The issue is not well understood, and quickly intractable.

Consequences can be dire.

The setting

We investigate the sensitivity of inferences to **influential sets**.

> *A set of observations $\mathcal{S}$ is influential if its omission has a large impact on some measure of interest λ when compared to others.*

The setting

We investigate the sensitivity of inferences to **influential sets**.

> *A set of observations $\mathcal{S}$ is influential if its omission has a large impact on some measure of interest λ when compared to others.*

We want the set with maximal influence $\Delta(\mathcal{S})$ on λ at given sizes — in order to find the **minimal influential set** $\mathcal{S}^{**}$, i.e. the *smallest set whose removal overturns a result of interest*.

The setting

We investigate the sensitivity of inferences to **influential sets**.

> *A set of observations $\mathcal{S}$ is influential if its omission has a large impact on some measure of interest λ when compared to others.*

We want the set with maximal influence $\Delta(\mathcal{S})$ on λ at given sizes — in order to find the **minimal influential set** $\mathcal{S}^{**}$, i.e. the *smallest set whose removal overturns a result of interest*.

Example — 'The Blessing of Bad Geography in Africa'

'[...] the differential effect of ruggedness is statistically significant and economically meaningful, [...]' (Nunn and Puga, 2012)

The issues — computation

Exactly determining the minimal influential set is usually *impossible*.

The issues — computation

Exactly determining the minimal influential set is usually *impossible*.

1. There are $\binom{N}{N_\alpha}$ potential sets, where $N_\alpha = |\mathcal{S}^{**}|$.
2. We need to compute λ, the quantity of interest, for each one.

The issues — computation

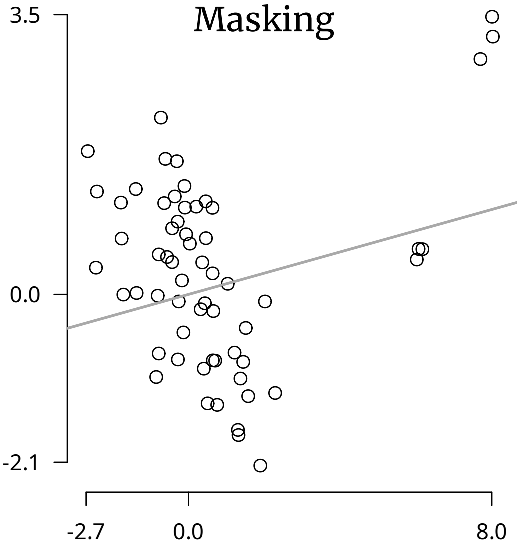
Exactly determining the minimal influential set is usually *impossible*.

1. There are $\binom{N}{N_\alpha}$ potential sets, where $N_\alpha = |\mathcal{S}^{**}|$.
2. We need to compute λ, the quantity of interest, for each one.

Consider $N = 1,000$, allowing for $N_\alpha = 10$, and assume that calculating λ takes one µs. Your sensitivity check will take about 8.35 billion years.

We rely on **approximations** in all but the simplest cases.

The issues — masking

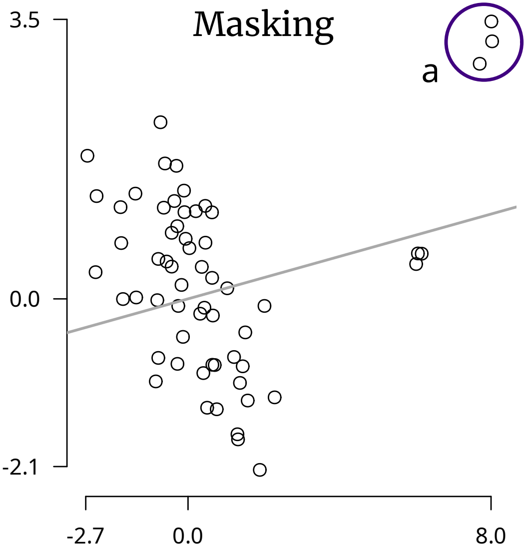


Consider the model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, with

$$\lambda(\mathcal{S}) = \left(\mathbf{X}'_{(\mathcal{S})}\mathbf{X}_{(\mathcal{S})}\right)^{-1}\mathbf{X}'_{(\mathcal{S})}\mathbf{y}_{(\mathcal{S})},$$

where $\mathcal{S}$ is a set of observations, and subscripts indicate removal.

The issues — masking



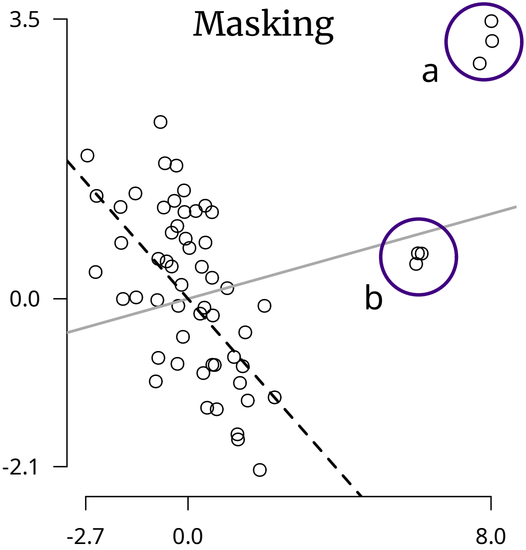
Consider the model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, with

$$\lambda(\mathcal{S}) = \left(\mathbf{X}'_{(\mathcal{S})}\mathbf{X}_{(\mathcal{S})}\right)^{-1}\mathbf{X}'_{(\mathcal{S})}\mathbf{y}_{(\mathcal{S})},$$

where $\mathcal{S}$ is a set of observations, and subscripts indicate removal.

- The set marked **'a'** is **highly influential** on the slope.

The issues — masking



Consider the model $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, with

$$\lambda(\mathcal{S}) = \left(\mathbf{X}'_{(\mathcal{S})}\mathbf{X}_{(\mathcal{S})}\right)^{-1}\mathbf{X}'_{(\mathcal{S})}\mathbf{y}_{(\mathcal{S})},$$

where $\mathcal{S}$ is a set of observations, and subscripts indicate removal.

- The set marked **'a'** is **highly influential** on the slope.
- However, it initially **masks** the influential set marked **'b'**.

Identifying influential sets

How do we identify a minimal influential set?

Identifying influential sets

How do we identify a minimal influential set?

We consider *three algorithms* to approximate $\mathcal{S}$ and $\Delta(\hat{\mathcal{S}})$, that are

- easy to implement,
- computationally tractable,
- differently trade speed for accuracy.

The algorithms — an initial approximation

Algorithm 0

Idea: Approximate $\mathcal{S}$ based on initial influence and Δ via summation.

The algorithms — an initial approximation

Algorithm 0

Idea: Approximate $\mathcal{S}$ based on initial influence and Δ via summation.

0. Compute $\Delta(\{i\})$ for each observation i, let $\hat{\mathcal{S}} \leftarrow \varnothing$.
1. Let $\hat{\mathcal{S}} \leftarrow \hat{\mathcal{S}} \cup \arg\max \Delta(\{j\})$, for $j \notin \hat{\mathcal{S}}$.
2. Let $\hat{\Delta}(\hat{\mathcal{S}}) \leftarrow \sum \Delta(\{k\})$ for all $k \in \hat{\mathcal{S}}$.
3. Go to step 1, unless $\hat{\Delta} > \Delta^*$ or $|\hat{\mathcal{S}}| > U$.

At $\mathcal{O}(1)$ complexity, **computing Δ dominates.** Broderick, Giordano, and Meager (2020) use a similar approach, approximating Δ ▸ Details.

The algorithms — divide and conquer

Algorithm 1

Idea: Approximate $\mathcal{S}$ based on initial influence; binary-search for Δ^*.

The algorithms — divide and conquer

Algorithm 1

Idea: Approximate $\mathcal{S}$ based on initial influence; binary-search for Δ^*.

1. Compute $\Delta(\{i\})$ for each observation i.
2. Create the ordered set $\mathcal{T}$ by ranking $\Delta(\{i\})$.
3. Binary-search for the smallest Δ^* in the interval (L, U).
   - Let $\hat{\mathcal{S}}$ be the first $(L+U)/2$ elements of $\mathcal{T}$.
   - Compute $\Delta(\hat{\mathcal{S}})$.
   - Adapt the lower or upper bound until done.

This adaptation yields improved precision at $\mathcal{O}(\log U)$ complexity.

The algorithms — an adaptive approximation

Algorithm 2

Idea: Adaptively and greedily build approximations to $\mathcal{S}$.

The algorithms — an adaptive approximation

Algorithm 2

Idea: Adaptively and greedily build approximations to $\mathcal{S}$.

0. Let $\hat{\mathcal{S}} \leftarrow \varnothing$.
1. Compute $\Delta(\hat{\mathcal{S}} \cup \{j\})$ for each $j \notin \hat{\mathcal{S}}$.
2. Let $\hat{\mathcal{S}} \leftarrow \hat{\mathcal{S}} \cup \arg\max \Delta(\hat{\mathcal{S}} \cup \{j\})$.
3. Go to step 1, unless $\Delta(\hat{\mathcal{S}}) > \Delta^*$ or $|\hat{\mathcal{S}}| > U$.

Now, we can *adapt for masking* at $\mathcal{O}(N_\alpha)$ complexity — computing Δ would still dominate, however, **efficient updating formulae** that facilitate computation are often available.

The quantity λ and computing Δ

Example — 'The Blessing of Bad Geography in Africa'

Rugged terrain hinders development globally. Nunn and Puga find a *different* (statistically and economically significant) *effect* in Africa.

The quantity λ and computing Δ

In most regression analyses, we tend to care about the

- an estimated **coefficient** ($\hat{\beta}$), and
- **uncertainty** around it (perhaps quantified via t values).

Example — 'The Blessing of Bad Geography in Africa'

Rugged terrain hinders development globally. Nunn and Puga find a *different* (statistically and economically significant) *effect* in Africa.

The quantity λ and computing Δ

In most regression analyses, we tend to care about the

- an estimated **coefficient** ($\hat{\beta}$), and
- **uncertainty** around it (perhaps quantified via t values).

Coefficient influence, e.g., is a function of errors and leverage, i.e.

$$\Delta(\{i\}) = \beta_{(\emptyset)} - \beta_{(\{i\})} = \frac{(\mathbf{X}'\mathbf{X})^{-1} x_i' e_i}{1 - h_i}.$$

Example — 'The Blessing of Bad Geography in Africa'

Rugged terrain hinders development globally. Nunn and Puga find a *different* (statistically and economically significant) *effect* in Africa.

A demonstration

*What does a **minimal influential set** look like in practice?*

A demonstration

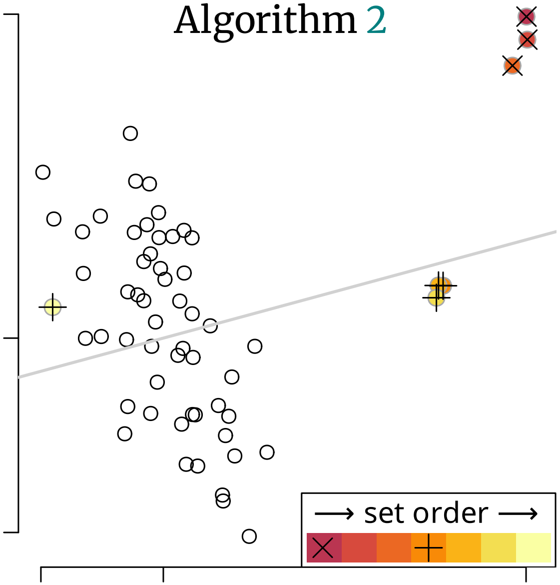
*What does a **minimal influential set** look like in practice?*

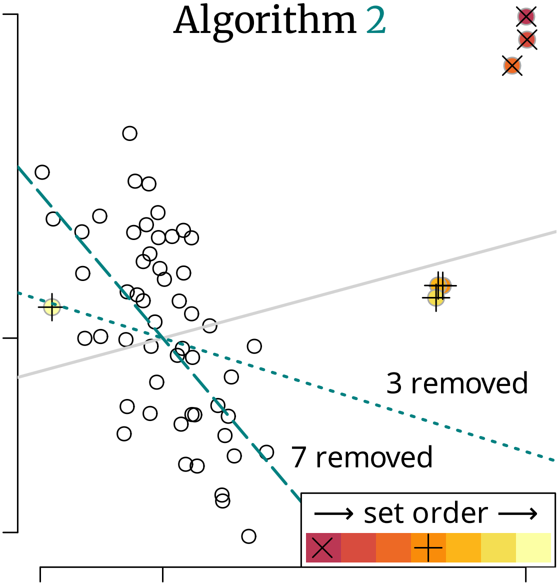
- We'll revisit the *univariate regression* to demonstrate.

A demonstration

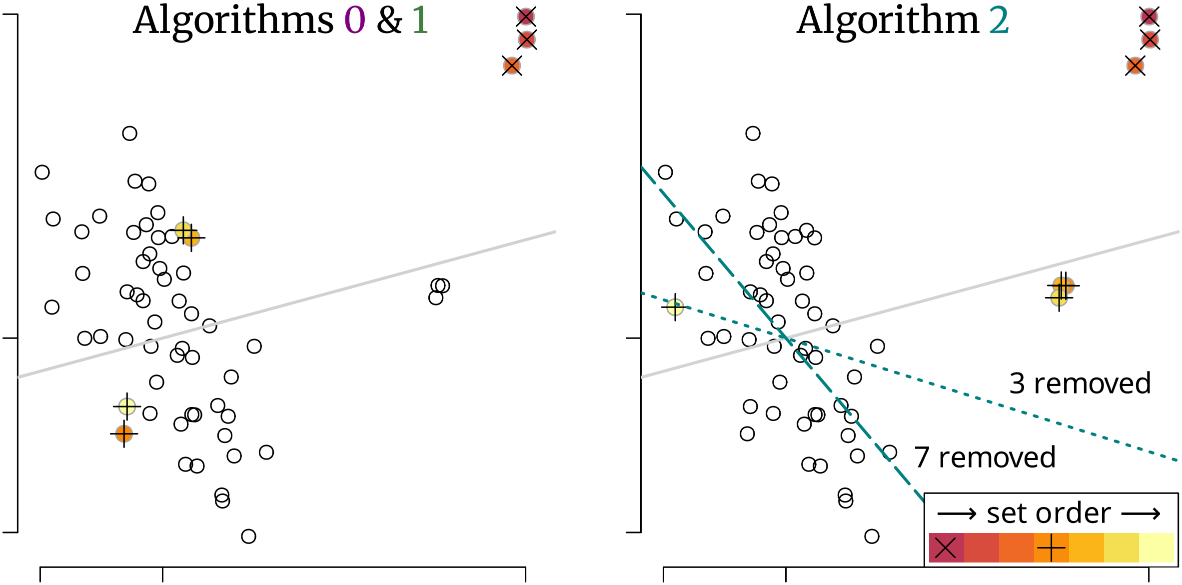
*What does a **minimal influential set** look like in practice?*

- We'll revisit the *univariate regression* to demonstrate.
- Then, we'll investigate *'The Blessing of **Bad Geography** in Africa'*.
    - Our target will be the t value of the main result.

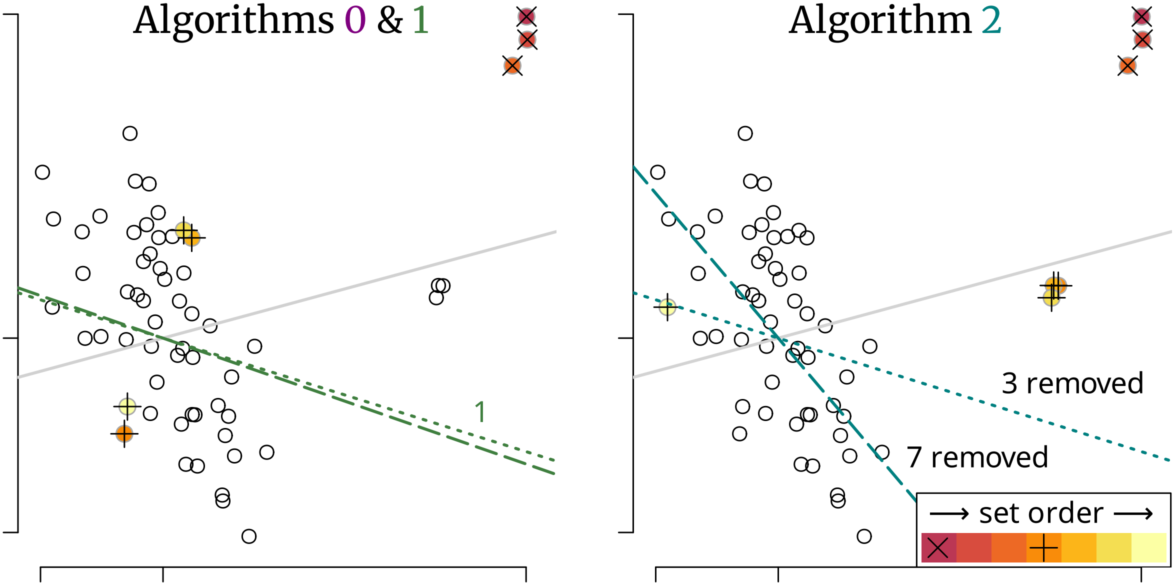

Algorithm 2
⟶ set order ⟶

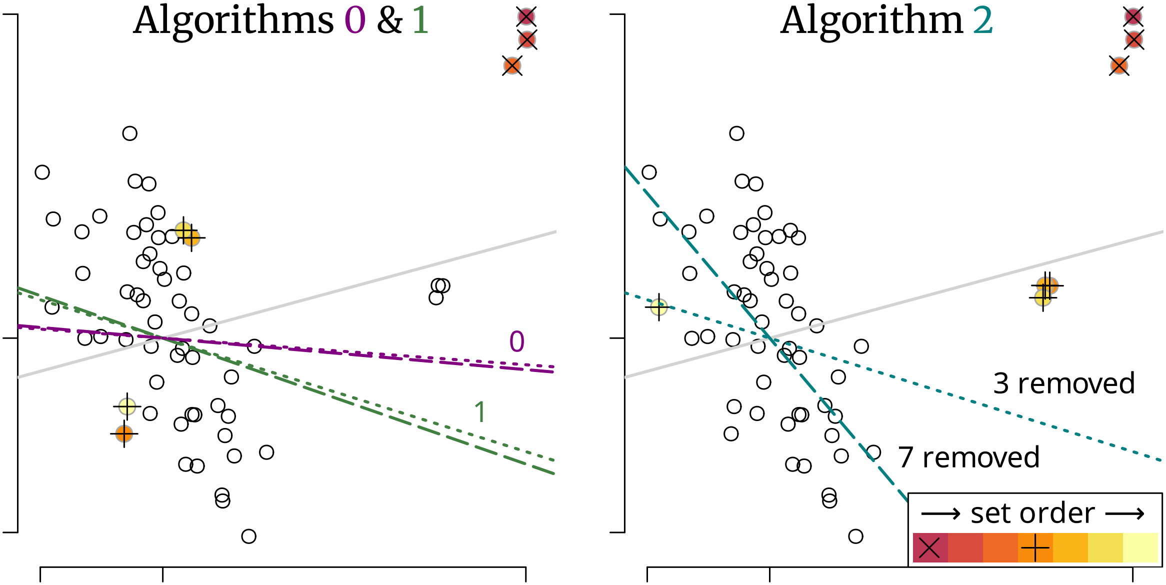


Algorithm 2
3 removed
7 removed
⟶ set order ⟶



Algorithms 0 & 1
Algorithm 2
3 removed
7 removed
⟶ set order ⟶


Algorithms 0 & 1
1
Algorithm 2
3 removed
7 removed
⟶ set order ⟶



Algorithms 0 & 1
0
1
Algorithm 2
3 removed
7 removed
⟶ set order ⟶

An application — influential sets and ruggedness

$\log$ GDP/capita $\sim$	Baseline	Plain
ruggedness, Africa†	0.321	0.302
	(2.53)	(2.32)
ruggedness	-0.231	-0.193
	(-2.99)	(-2.38)
coast distance	Yes	Yes
other controls	Yes	–
observations	170	170

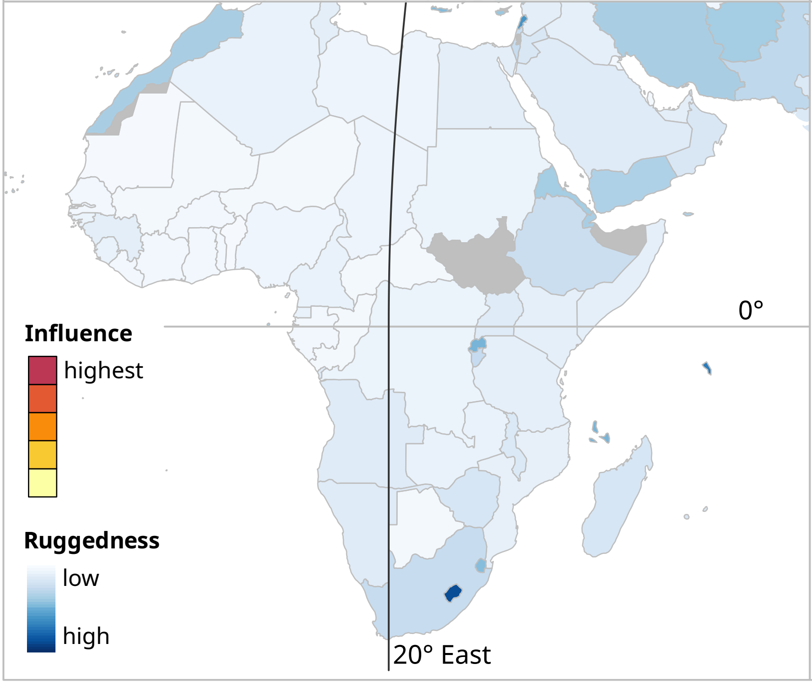
The (t values) are based on HC1 standard errors. The 'thresholds' indicate the number of removed observation that nullify significance (at the 5% level), [flip the sign], and {significantly flip the sign}.

An application — influential sets and ruggedness

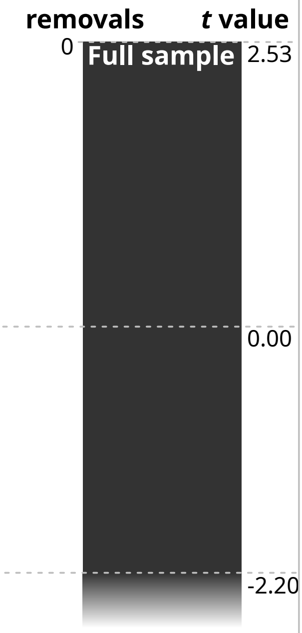
log GDP/capita ~	Baseline	Plain
ruggedness, Africa†	0.321	0.302
	(2.53)	(2.32)
ruggedness	-0.231	-0.193
	(-2.99)	(-2.38)
coast distance	Yes	Yes
other controls	Yes	–
observations	170	170
thresholds†	2 [5]{11}	2[7]{16}

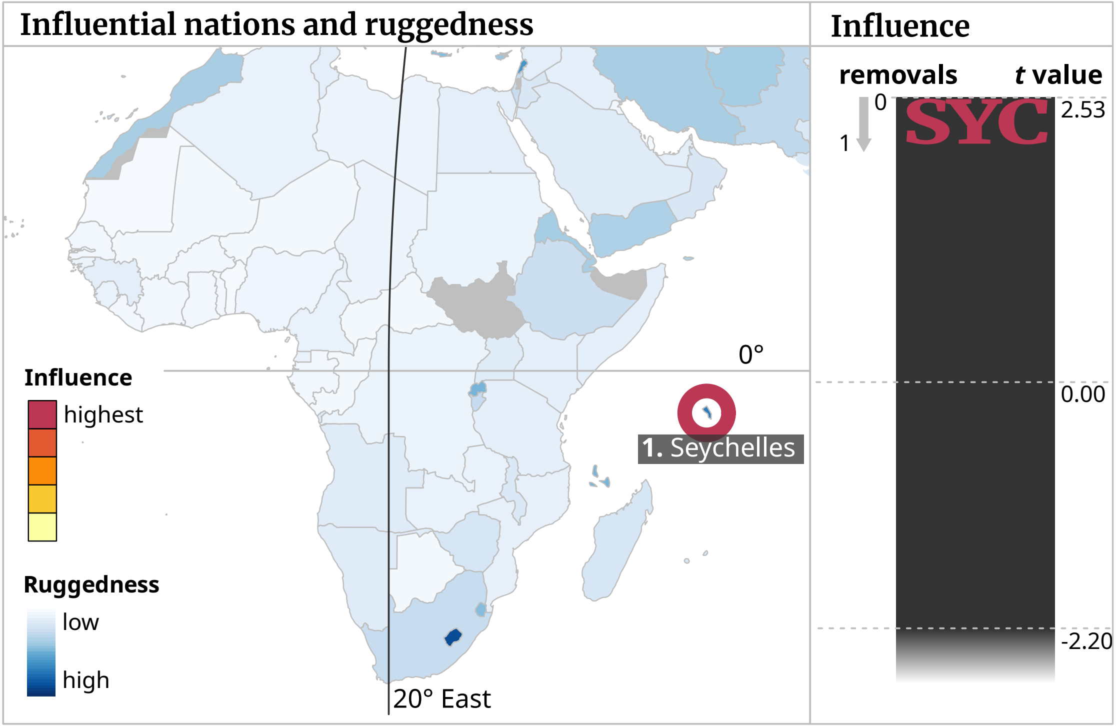
The (t values) are based on HC1 standard errors. The 'thresholds' indicate the number of removed observation that nullify significance (at the 5% level), [flip the sign], and {significantly flip the sign}.

Influential nations and ruggedness

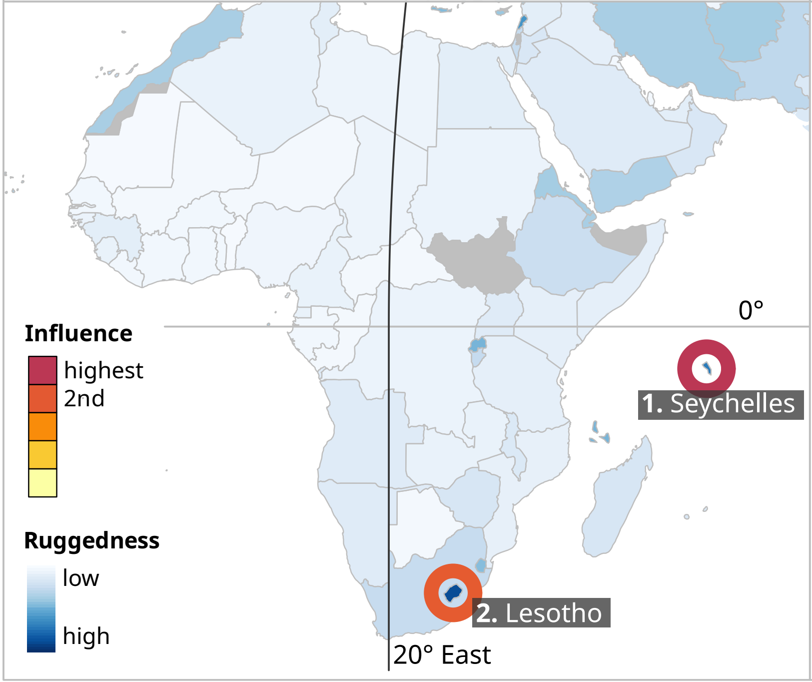


Influence

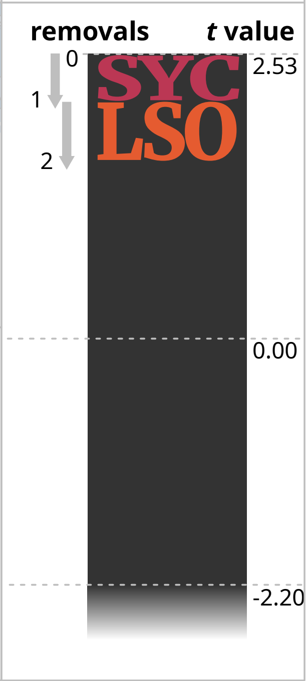



Influential nations and ruggedness
Influence
removals
t value
0
1
SYC
2.53
0.00
-2.20
0°
1. Seychelles
Influence
highest
Ruggedness
low
high
20° East

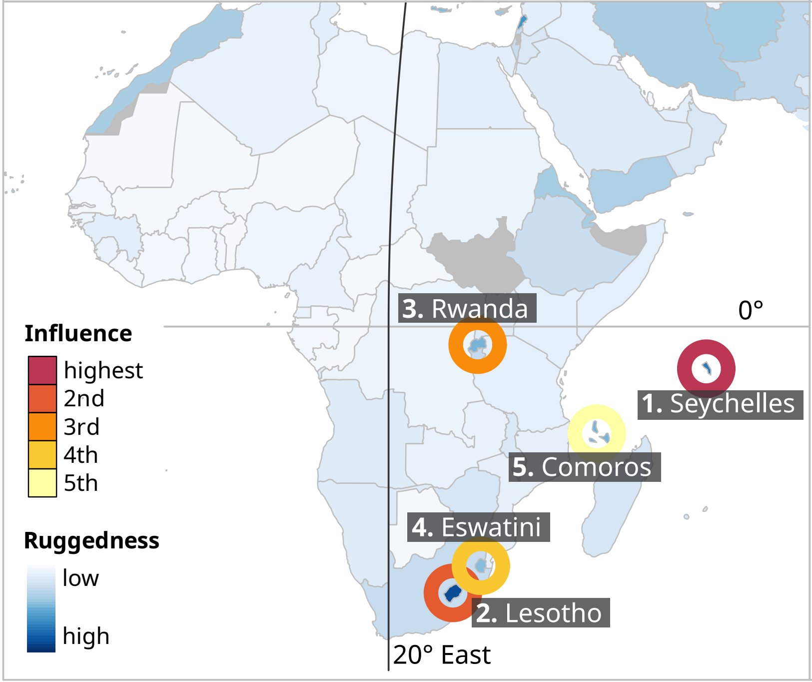
Influential nations and ruggedness



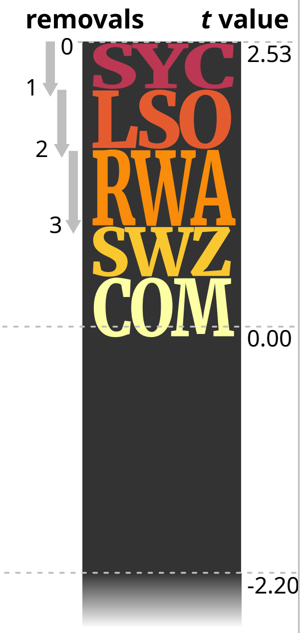
Influence



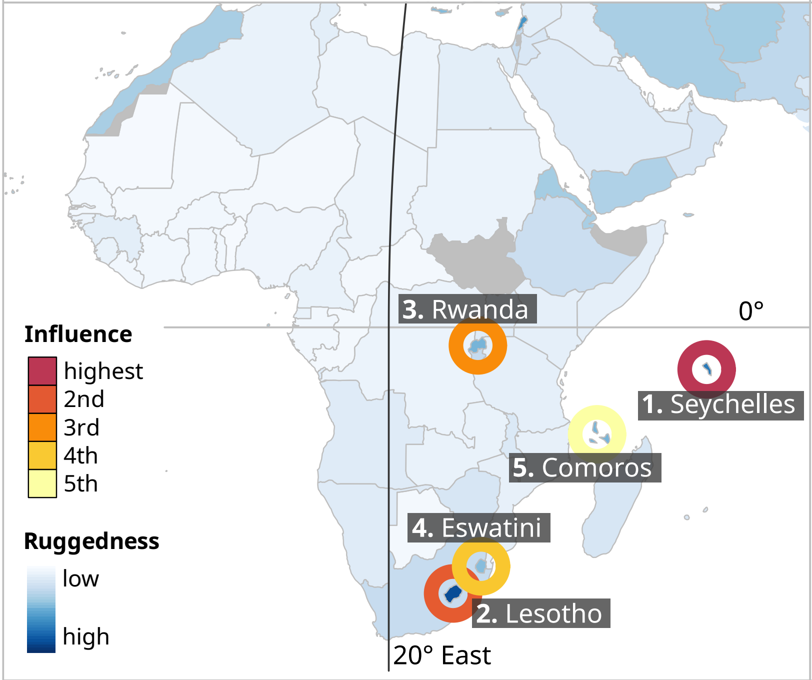
Influential nations and ruggedness



Influence



Influential nations and ruggedness



Influence

Summary

To wrap up — we were looking for the **minimal influential set**, i.e. an

Summary

To wrap up — we were looking for the **minimal influential set**, i.e. an

- **intuitive** (*two nations remove significance*),
- **insightful** (*confounders, heterogeneity, validity*),
- **widely applicable** (*size, clustered errors, 2SLS*) sensitivity check.

Summary

To wrap up — we were looking for the **minimal influential set**, i.e. an

- intuitive (*two nations remove significance*),
- insightful (*confounders, heterogeneity, validity*),
- widely applicable (*size, clustered errors, 2SLS*) sensitivity check.

We have also caused some issues …

- What does sensitivity **imply**? ‣ My two cents
- How to find **better sets faster**?

Summary

To wrap up — we were looking for the **minimal influential set**, i.e. an

- intuitive (*two nations remove significance*),
- insightful (*confounders, heterogeneity, validity*),
- widely applicable (*size, clustered errors, 2SLS*) sensitivity check.

We have also caused some issues ...

- What does sensitivity **imply**? My two cents
- How to find **better sets faster**?

Find me & the paper.

References i

Anthony C. Atkinson, Marco Riani, and Andrea Cerioli.
The forward search: theory and data analysis.
Journal of the Korean Statistical Society, 39(2):117–134, 2010.

Tamara Broderick, Ryan Giordano, and Rachael Meager.
An automatic finite-sample robustness metric: can dropping a little data change conclusions?, 2020.

Jesús Crespo Cuaresma, Stephan Klasen, and Konstantin M. Wacker.
When do we see poverty convergence?
Oxford Bulletin of Economics and Statistics, 2022.

Bradley Efron and Robert J. Tibshirani.
***An introduction to the Bootstrap*.**
CRC Press, 1994.

Ryan Giordano, Runjing Liu, Michael I. Jordan, and Tamara Broderick.
Evaluating sensitivity to the stick-breaking prior in Bayesian nonparametrics.
Bayesian Analysis, -1(-1):1–34, 2022.

Nathan Nunn and Diego Puga.
Ruggedness: the blessing of bad geography in Africa.
Review of Economics and Statistics, 94(1):20–36, 2012.

Broderick, Giordano and Meager (2020)

'Can Dropping a Little Data Change Conclusions?' — the authors check using the 'Approximate Maximum Influence Perturbation'.

- Computation is effectively instant.
  - Their algorithm is a special case of Algorithm 0.
  - They use a linear approximation to compute Δ.
- **Accuracy suffers especially when influential sets are present**.
  - *Masking* issues and *downward bias*, akin to Algorithm 0.
  - Their approximation of e.g. $\beta_{(\varnothing)} - \beta_{(\{i\})}$ *discards the leverage*, whereas

$$\text{influence} = f(\text{errors}, \text{leverage}).$$

- An option for settings with non-tractable Δ (see Giordano, 2022).

Go back

Microcredit — seven randomised control trials

Sensitivity of the average treatment effect of microcredits

study region	BIH		MON		ETH		MEX		MOR		PHI		IND	
algorithm	(0)	(2)	(0)	(2)	(0)	(2)	(0)	(2)	(0)	(2)	(0)	(2)	(0)	(2)
sign-switch	14	13	16	15	1	1	1	1	11	11	9	9	6	6
significance	49	39	43	37	117	13	20	12	35	33	74	54	41	35
observations	1,195		961		3,113		16,560		5,498		1,113		6,863	

The reported values are the number of removals needed to induce a sign-switch of the average treatment effect, and have this sign-flipped coefficient become significant (at the 1% level) using Algorithm 0 and 2. Algorithm 2 outperforms consistently, but few observations are needed to overturn results in all cases.

Learning from influential sets — ruggedness

$\log$ GDP/capita $\sim$	Baseline	Plain	Population	Area
ruggedness, Africa†	0.321	0.302	0.190	0.215
	(2.53)	(2.32)	(1.66)	(1.63)
ruggedness	-0.231	-0.193	-0.231	-0.238
	(-2.99)	(-2.38)	(-2.94)	(-3.08)
coast distance	Yes	Yes	Yes	Yes
population in 1400	–	–	Yes	–
land area	–	–	–	Yes
other controls	Yes	–	Yes	Yes
observations	170	170	168	170
thresholds†	2[5]{11}	2[7]{16}	–[3]{6}	–[4]{8}

The 'thresholds' indicate the number of removed observation that nullify significance (at the 5% level), [flip the sign], and {significantly flip the sign}. The t values in (brackets) are based on HC1 errors. ▸ Go back

Implications

A result **seems too sensitive** due to a **small** minimal influential set ...

- We are searching for the **needle in the haystack**.
  - \+ Small in relative terms should be fine.
  - \- Small in absolute terms indicates low power.
- We are not — there should be plenty of needles.
  - ! We have a **classical outlier problem**, and some data to investigate.
  - ? Are there unobserved confounders, heterogeneous effects, etc.

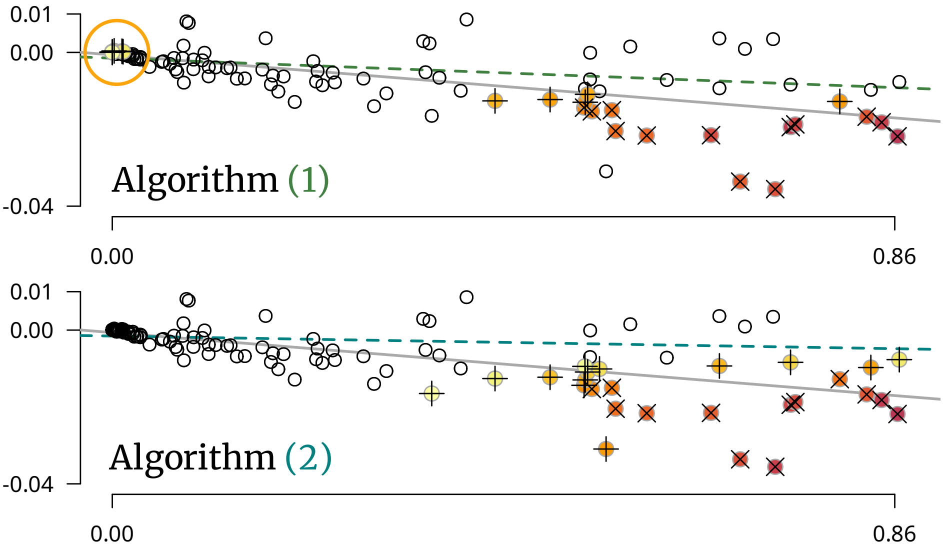
In any case, this is a prompt to use more comprehensive measures, e.g. forward-search by Atkinson, Riani, and Cerioli (2010), or Bootstrap methods. Go back

The origins of mistrust

	Trust of relatives ~		Trust of neighbours ~	
	Pooled	West | East	Pooled	West | East
exports/area†	-0.133	-0.145	-0.159	-0.168
	(-3.68)	(-3.84)	(-4.67)	(-4.48)
exports/area, East		0.053		0.023
		(0.96)		(0.32)
individual controls	Yes	Yes	Yes	Yes
district controls	Yes	Yes	Yes	Yes
country fixed effects	Yes	Yes	Yes	Yes
observations	20,062	7,549 | 12,513	20,027	7,523 | 12,504
thresholds†	105[380]{656}	78[301]{532}	161[425]{768}	133[323]{527}
ethnicity clusters	185	62 | 123	185	62 | 123
district clusters	1,257	628 | 651	1,257	628 | 651

The (t values) are based on 2-way clustered standard errors. The ‘thresholds’ indicate the number of removed observation that nullify significance (at the 1% level), [flip the sign], and {significantly do so}.

Poverty convergence



Data and regression line for the poverty convergence regression of Crespo Cuaresma et al. (2022), before (solid line) and after (dashed line) removing the influential set $\hat{\mathcal{S}}^*_{26}$. There are 126 observations in total.