Forecasting the $Y_{500} - M_{500}$ scaling relation from the NIKA2 SZ Large Program

Florian Kéruzoré, LPSC, Grenoble

mmUniverse@NIKA2, June 2021

financé par
IDEX Université Grenoble Alpes

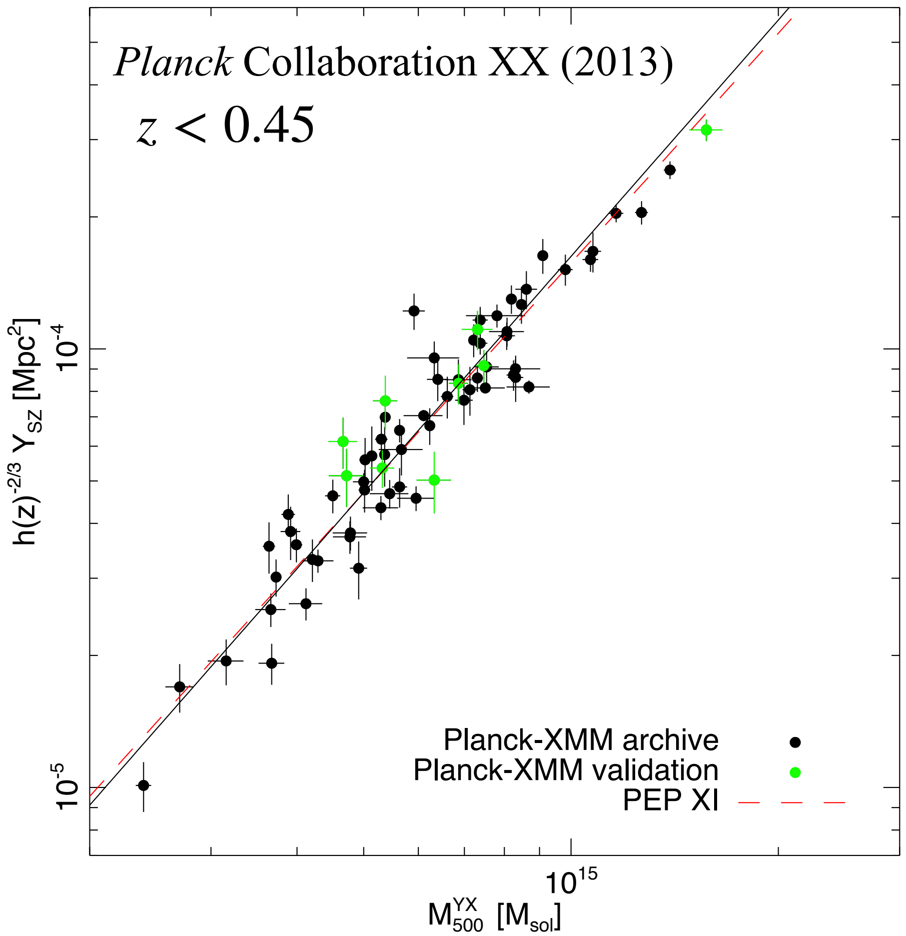
Introduction

Florian Kéruzoré
mmUniverse
June 2021

- Cluster masses are needed for cosmology, but not a direct observable

  → Empirical mass-observable scaling relations **(SR)** are calibrated on small cluster samples

- One of the goals of the NIKA2 SZ Large Program (LPSZ — talk by L. Perotto):
  - SR between mass M_{500} & integrated Compton parameter Y_{500} (SZ survey observable)
  - Benefiting from NIKA2's high angular resolution: **better-constrained quantities**



→ Improvement over *Planck* measurement

③ Scope

- **This work:** preparing the measurement of the scaling relation from the LPSZ data
  - Setup a Bayesian hierarchical model regression scheme
  - Generate **mock LPSZ-like cluster samples**
  - Search for **biases** in the results, *i.e.* see how LPSZ data features affect the analysis
  - Begin **forecasting precision** given the sample size / data quality

④ Outline

Scaling relation adjustment

Realistic mock sample generation

Results: biases & precision

Conclusions

⑤ Notations & linear mass-observable relation

- Self-similar scenario of structure growth: **power law relation** between
  - integrated Compton parameter $D_{\rm A}^2\, Y_{500} \propto \int_0^{R_{500}} P_{\rm e}(r)\, r^2 \mathrm{d}r$
  - mass M_{500}

$$E^{-2/3}(z) \frac{D_{\rm A}^2\, Y_{500}}{10^{-4}\ \mathrm{Mpc}^2} = 10^{\alpha} \left[\frac{M_{500}}{6 \times 10^{14}\ \mathrm{M}_{\odot}}\right]^{\beta}$$

- Defining the log-scaled SZ observable Y and mass Z makes the scaling relation **linear:**

$$\left.\begin{aligned} Y &\equiv \log\left[E^{-2/3}(z) \frac{D_{\rm A}^2\, Y_{500}}{10^{-4}\ \mathrm{Mpc}^2}\right] \\ Z &\equiv \log\left[\frac{M_{500}}{6 \times 10^{14}\ \mathrm{M}_{\odot}}\right] \end{aligned}\right\} \Rightarrow \; Y = \alpha_{Y|Z} + \beta_{Y|Z} Z$$

- SR = trend: **intrinsic scatter** due to cluster physics → Gaussian scatter around the relation:

$$P(Y\,|\,Z) = \mathcal{N}(\alpha_{Y|Z} + \beta_{Y|Z} Z,\ \sigma^2_{Y|Z})$$

→ parameters of interest: $\alpha_{Y|Z}$ (intercept), $\beta_{Y|Z}$ (slope), $\sigma_{Y|Z}$ (intrinsic scatter)

Hierarchical modeling

- Bayesian hierarchical modeling of the SR (Kelly07, Andreon+13, Mantz15, Sereno16, …)
- **Gaussian intrinsic scatter** around the relation:

$$P(Y\,|\,Z) = \mathcal{N}(\alpha_{Y|Z} + \beta_{Y|Z} Z,\, \sigma^2_{Y|Z})$$

⑥ Hierarchical modeling

- Bayesian hierarchical modeling of the SR (Kelly07, Andreon+13, Mantz15, Sereno16, …)

- **Gaussian intrinsic scatter** around the relation:

$$P(Y\,|\,Z) = \mathcal{N}(\alpha_{Y|Z} + \beta_{Y|Z}Z,\ \sigma^2_{Y|Z})$$

- **Eddington bias:** We don't know the true mass Z, but a mass estimator X

$$P(X\,|\,Z) = \mathcal{N}(\alpha_{X|Z} + \beta_{X|Z}Z,\ \sigma^2_{X|Z})$$

⑥ Hierarchical modeling

- Bayesian hierarchical modeling of the SR (Kelly07, Andreon+13, Mantz15, Sereno16, …)

- **Gaussian intrinsic scatter** around the relation:

  $$P(Y\,|\,Z) = \mathcal{N}(\alpha_{Y|Z} + \beta_{Y|Z} Z,\, \sigma^2_{Y|Z})$$

- **Eddington bias:** We don't know the true mass Z, but a mass estimator X

  $$P(X\,|\,Z) = \mathcal{N}(\alpha_{X|Z} + \beta_{X|Z} Z,\, \sigma^2_{X|Z})$$

- **Measured values** (y, x) and uncertainties with covariance V:

  for each data point i, $\quad P(\{y_i,\, x_i\}\,|\,\{Y_i,\, X_i\}) = \mathcal{N}_2(\{Y_i,\, X_i\},\, V_i)$

⑥ Hierarchical modeling

- Bayesian hierarchical modeling of the SR (Kelly07, Andreon+13, Mantz15, Sereno16, …)

- **Gaussian intrinsic scatter** around the relation:

  $P(Y\,|\,Z) = \mathcal{N}(\alpha_{Y|Z} + \beta_{Y|Z} Z,\, \sigma^2_{Y|Z})$

- **Eddington bias:** We don't know the true mass Z, but a mass estimator X

  $P(X\,|\,Z) = \mathcal{N}(\alpha_{X|Z} + \beta_{X|Z} Z,\, \sigma^2_{X|Z})$

- **Measured values** (y, x) and uncertainties with covariance V:

  for each data point i, $P(\{y_i, x_i\}\,|\,\{Y_i, X_i\}) = \mathcal{N}_2(\{Y_i, X_i\},\, V_i)$

- **Malmquist bias (MB):** only objects above $y_{\rm th}$ are detectable

  for each data point i, $P(\{y_i, x_i\}\,|\,\{Y_i, X_i\}) \propto \mathcal{N}_2(\{Y_i, X_i\},\, V_i) \times H(y_i - y_{\rm th})$

  (truncated probability distribution)

$H(x)$ = Heaviside step function

6 Hierarchical modeling

- Bayesian hierarchical modeling of the SR (Kelly07, Andreon+13, Mantz15, Sereno16, …)

- **Gaussian intrinsic scatter** around the relation:

$$P(Y\,|\,Z) = \mathcal{N}(\alpha_{Y|Z} + \beta_{Y|Z}Z,\, \sigma^2_{Y|Z})$$

- **Eddington bias:** We don't know the true mass Z, but a mass estimator X

$$P(X\,|\,Z) = \mathcal{N}(\alpha_{X|Z} + \beta_{X|Z}Z,\, \sigma^2_{X|Z})$$

- **Measured values** (y, x) and uncertainties with covariance V:

for each data point i, $\quad P(\{y_i, x_i\}\,|\,\{Y_i, X_i\}) = \mathcal{N}_2(\{Y_i, X_i\}, V_i)$

- **Malmquist bias (MB):** only objects above $y_{\rm th}$ are detectable

for each data point i, $\quad P(\{y_i, x_i\}\,|\,\{Y_i, X_i\}) \propto \mathcal{N}_2(\{Y_i, X_i\}, V_i) \times H(y_i - y_{\rm th})$

(truncated probability distribution)

$H(x)$ = Heaviside step function

- **Effect of latent distribution:** model the intrinsic distribution of the mass as a gaussian mixture

$$P(Z) = (1/n_{\rm mix}) \sum_{k=1}^{n_{\rm mix}} \pi_k \mathcal{N}(\mu_k,\, \sigma_k^2)$$

⑥ Hierarchical modeling

Florian Kéruzoré
mmUniverse
June 2021

Inputs
Parameters of interest
Nuisance parameters

- Bayesian hierarchical modeling of the SR (Kelly07, Andreon+13, Mantz15, Sereno16, …)
- **Gaussian intrinsic scatter** around the relation:

$$P(Y|Z) = \mathcal{N}(\alpha_{Y|Z} + \beta_{Y|Z} Z, \sigma^2_{Y|Z})$$

- **Eddington bias:** We don't know the true mass Z, but a mass estimator X

$$P(X|Z) = \mathcal{N}(\alpha_{X|Z} + \beta_{X|Z} Z, \sigma^2_{X|Z})$$

- **Measured values** (y, x) and uncertainties with covariance V:

for each data point i, $\quad P(\{y_i, x_i\} \,|\, \{Y_i, X_i\}) = \mathcal{N}_2(\{Y_i, X_i\}, V_i)$

- **Malmquist bias (MB):** only objects above y_{th} are detectable

for each data point i, $\quad P(\{y_i, x_i\} \,|\, \{Y_i, X_i\}) \propto \mathcal{N}_2(\{Y_i, X_i\}, V_i) \times H(y_i - y_{\mathrm{th}})$

(truncated probability distribution)

$H(x)$ = Heaviside step function

- **Effect of latent distribution:** model the intrinsic distribution of the mass as a gaussian mixture

$$P(Z) = (1/n_{\mathrm{mix}}) \sum_{k=1}^{n_{\mathrm{mix}}} \pi_k \mathcal{N}(\mu_k, \sigma^2_k)$$

⑦ MCMC sampling using LIRA

- Including priors on the parameters gives the **posterior distribution** $P(\vartheta \,|\, \{x_i, y_i\})$ to be sampled

- We use the `LIRA` library in `R` (Sereno16)
  - LInear Regression in Astronomy → designed to take into account common astronomical data features
  - And even more: can take into account several other features
    Linearity break, mass-dependent scatter, redshift evolution, …
  - Uses the hierarchical model described in previous slide
  - Uses Gibbs sampling MCMC to perform the regression
    very well-suited to high-dimensional bayesian hierarchical models
  - Well documented, validated on simulated datasets (arXiv:1509.05778)

- `LIRA` used to sample the posterior distribution in the parameter space
  With uninformative priors on parameters

Outline

Scaling relation adjustment

Realistic mock sample generation

Results: biases & precision

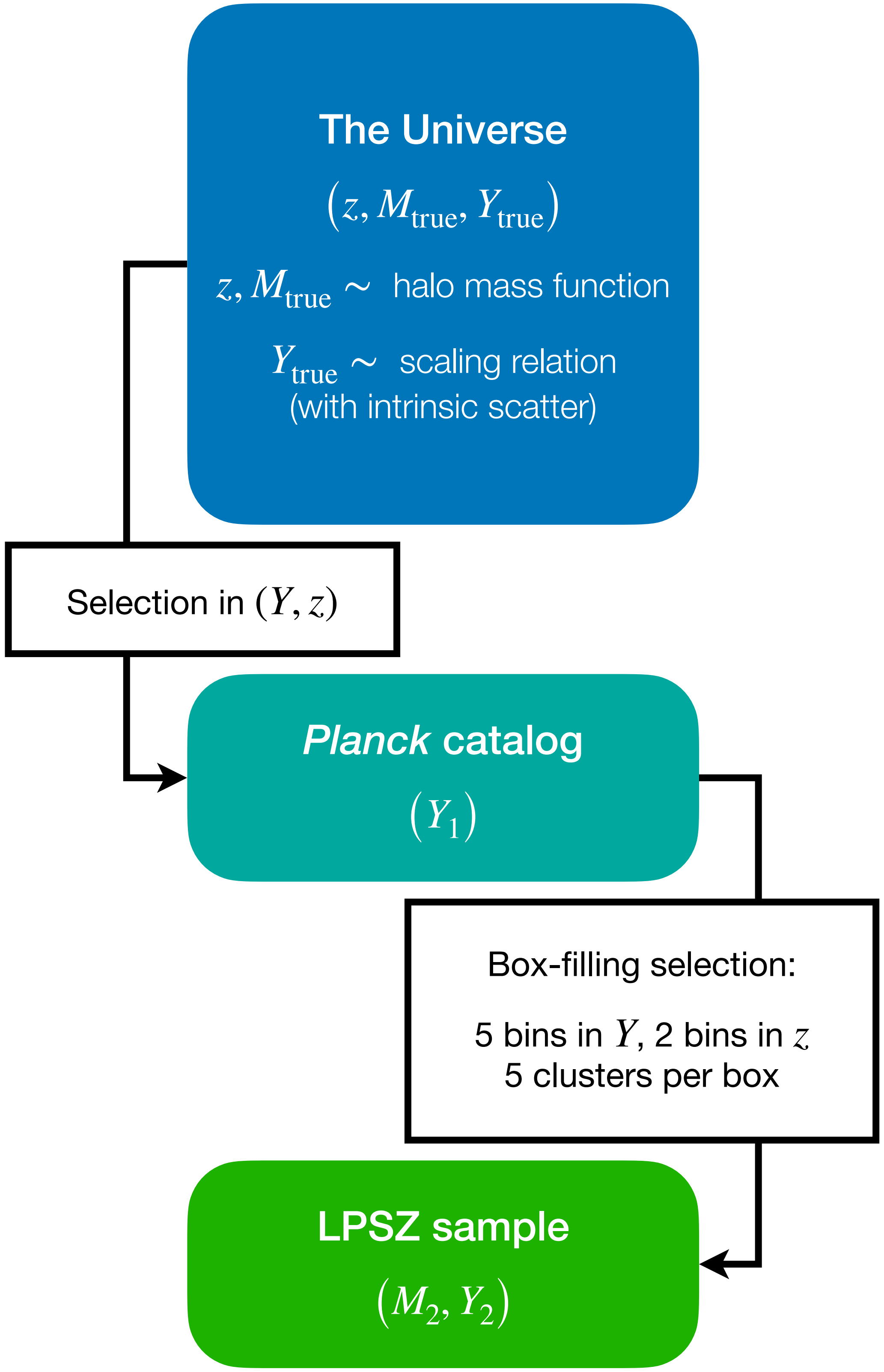
Conclusions

9 LPSZ mock sample generation

- **Goal:** what can we expect from the NIKA2 LPSZ for SR?
  - Generate “mock” cluster samples that **mimic the LPSZ**, with known SR
  - Fit them using the model presented before
  - Test the analysis’ **accuracy** (check for biases) and **precision** (evaluate uncertainties)

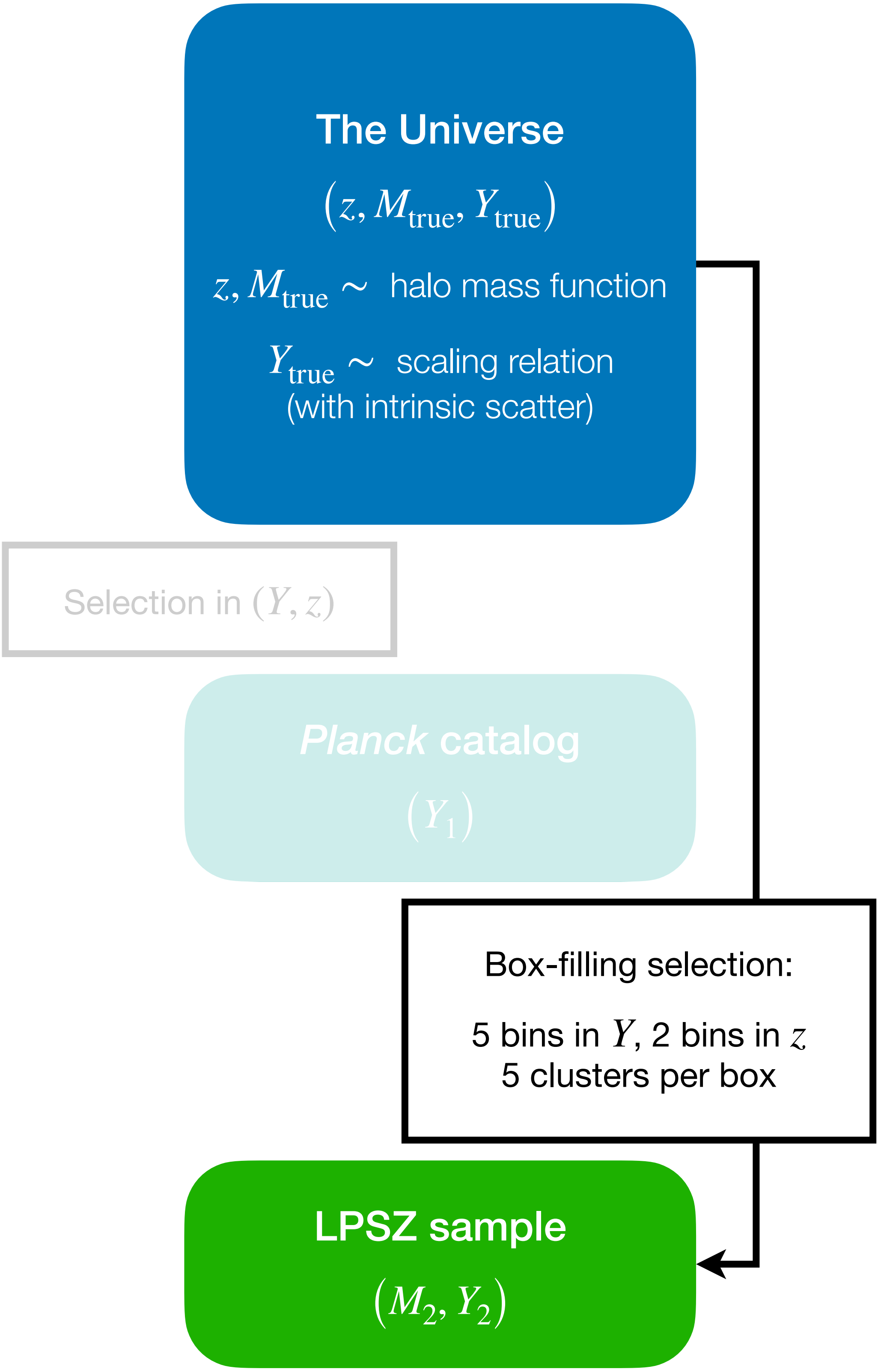
⑨ LPSZ mock sample generation

- **Goal:** what can we expect from the NIKA2 LPSZ for SR?
  - Generate “mock” cluster samples that **mimic the LPSZ**, with known SR
  - Fit them using the model presented before
  - Test the analysis’ **accuracy** (check for biases) and **precision** (evaluate uncertainties)
- **LPSZ selection function:**
  - 10 “boxes” from 2 bins in z and 5 in Y
  - Fill boxes with clusters from *Planck*/ACT catalogs
  - Measure (Y, M) for each cluster using SZ+Xrays



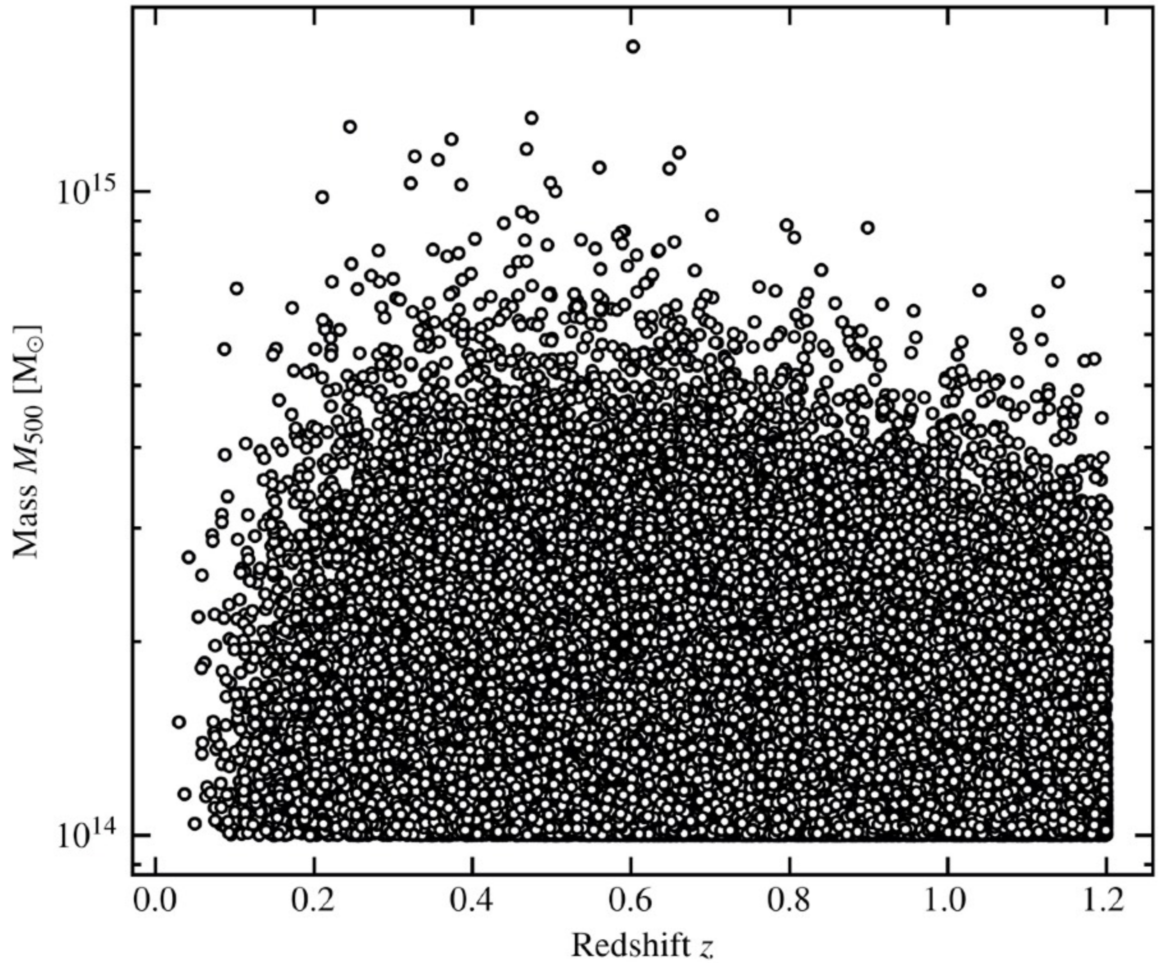
LPSZ mock sample generation

- **Goal:** what can we expect from the NIKA2 LPSZ for SR?
  - Generate "mock" cluster samples that **mimic the LPSZ**, with known SR
  - Fit them using the model presented before
  - Test the analysis' **accuracy** (check for biases) and **precision** (evaluate uncertainties)
- **LPSZ selection function:**
  - 10 "boxes" from 2 bins in z and 5 in Y
  - Fill boxes with clusters from *Planck*/ACT catalogs
  - Measure (Y, M) for each cluster using SZ+Xrays
- **This work:**
  - Create random LPSZ-like samples
  - Bypass *Planck*+ACT selection step
    → Ignore their selection function



LPSZ mock sample generation

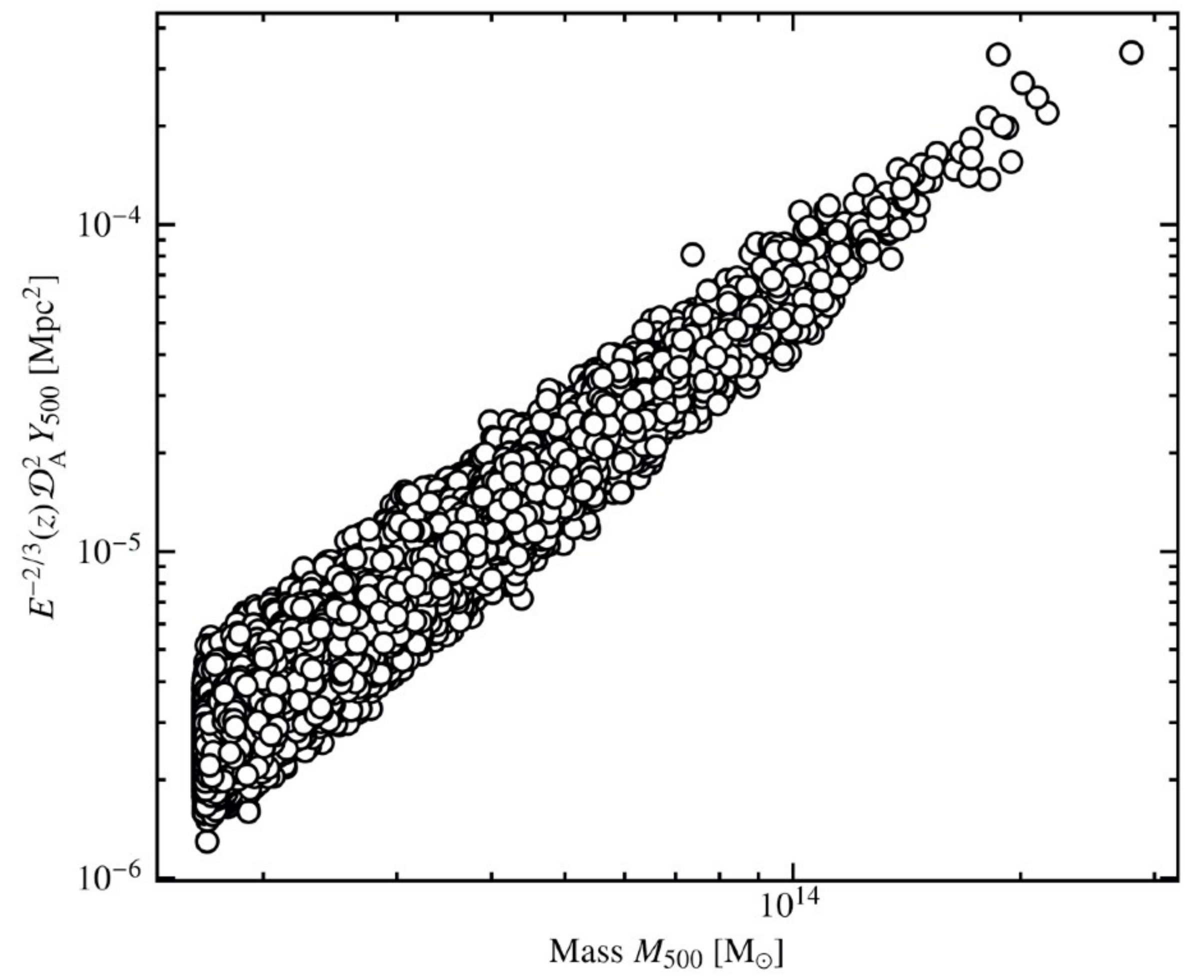
- **Step 1:** draw random (z, M_{500}) points
  from a Tinker+08 halo mass function



LPSZ mock sample generation

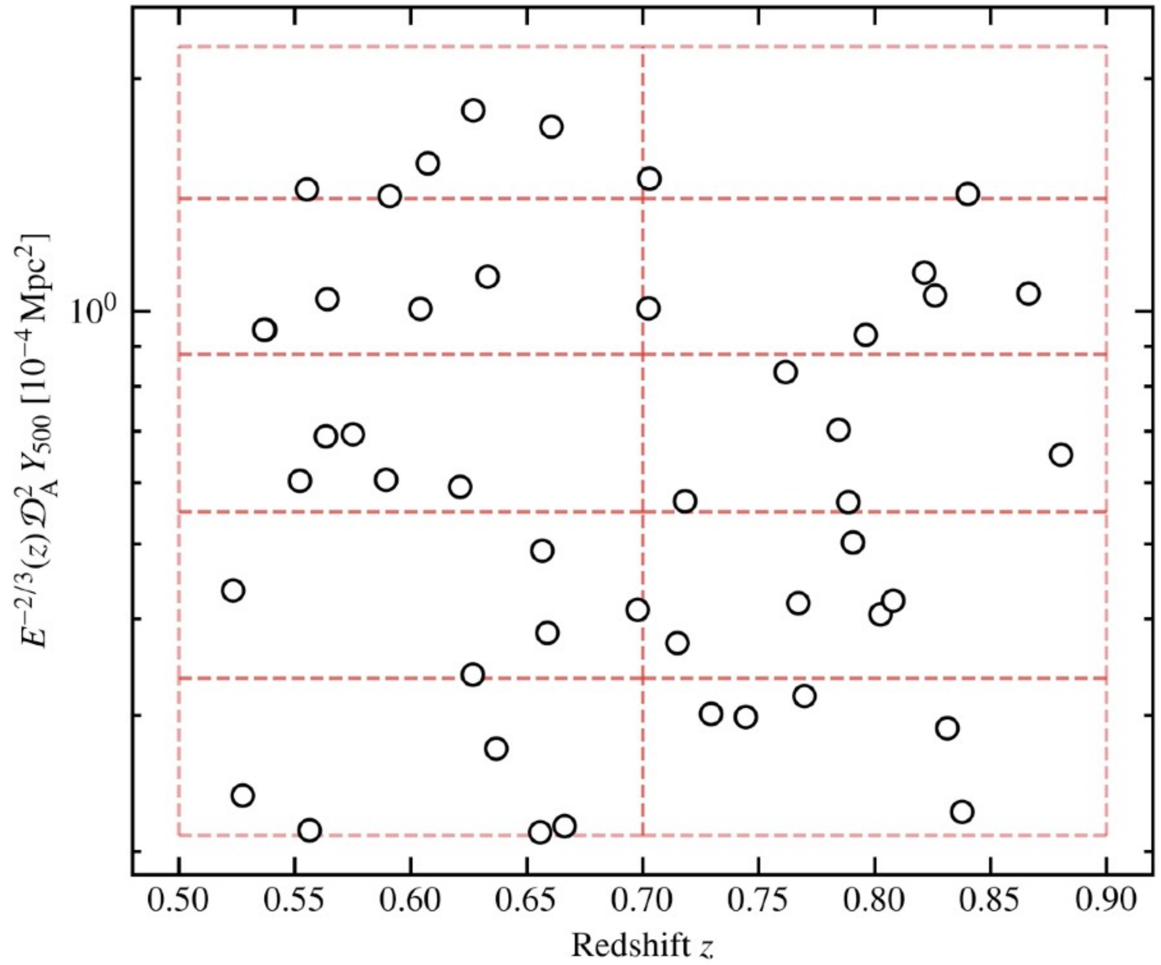
- **Step 1:** draw random (z, M_{500}) points
  from a Tinker+08 halo mass function

- **Step 2:** apply fiducial input SR → observable Y values
  - *Planck* results as truth:
    $\alpha_{Y|Z} = -0.19, \beta_{Y|Z} = 1.79, \sigma_{Y|Z} = 0.075$



⑩ LPSZ mock sample generation

- **Step 1:** draw random (z, M_{500}) points
  from a Tinker+08 halo mass function
- **Step 2:** apply fiducial input SR → observable Y values
  - *Planck* results as truth:
    $\alpha_{Y|Z} = -0.19,\ \beta_{Y|Z} = 1.79,\ \sigma_{Y|Z} = 0.075$
- **Step 3:** apply the box-filling algorithm to select an LPSZ-like sample
  - Same (Y, z) bins as the real LPSZ
  - 5 clusters/box → 50 total (~real LPSZ)



⑩ LPSZ mock sample generation

Florian Kéruzoré
mmUniverse
June 2021

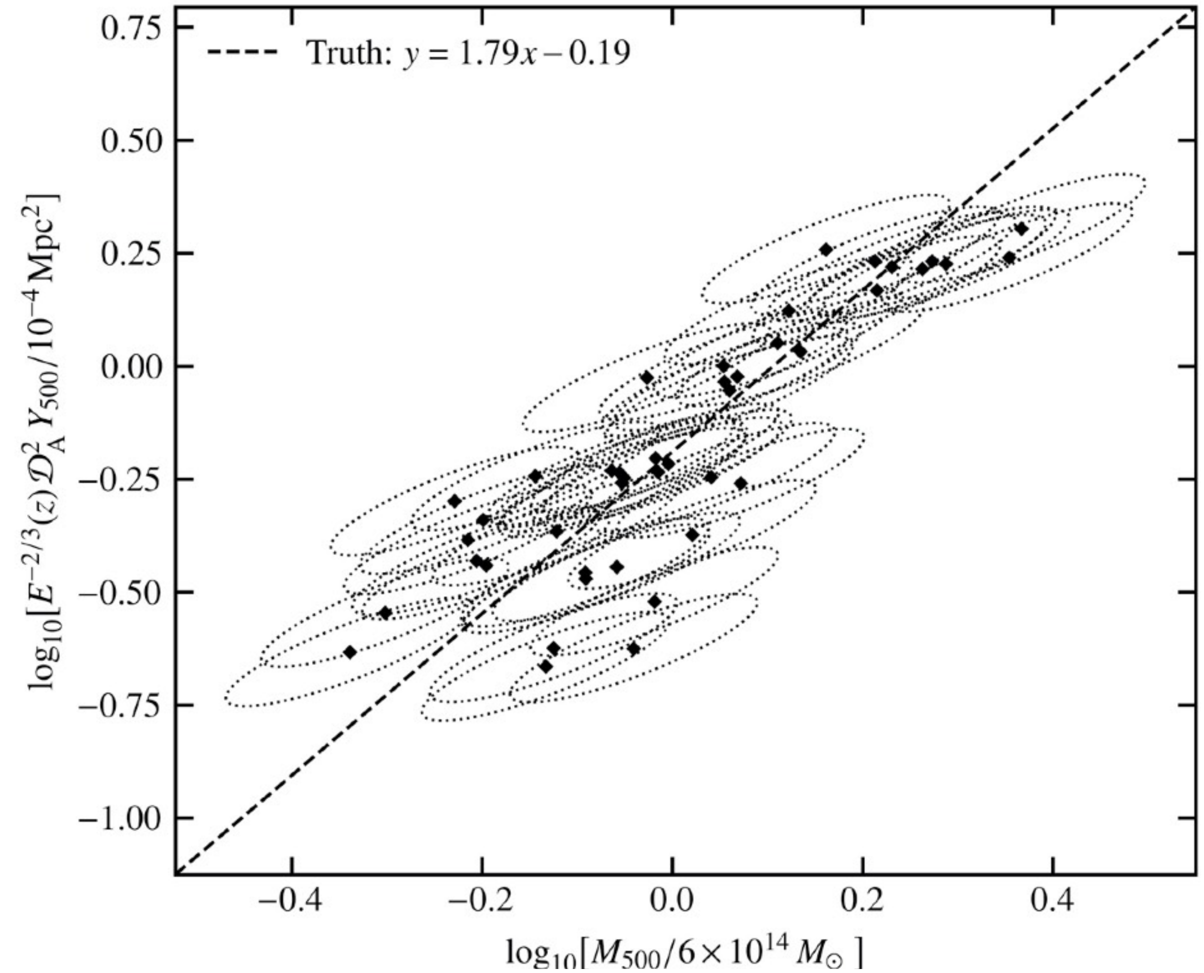
- **Step 1:** draw random (z, M_{500}) points
  from a Tinker+08 halo mass function

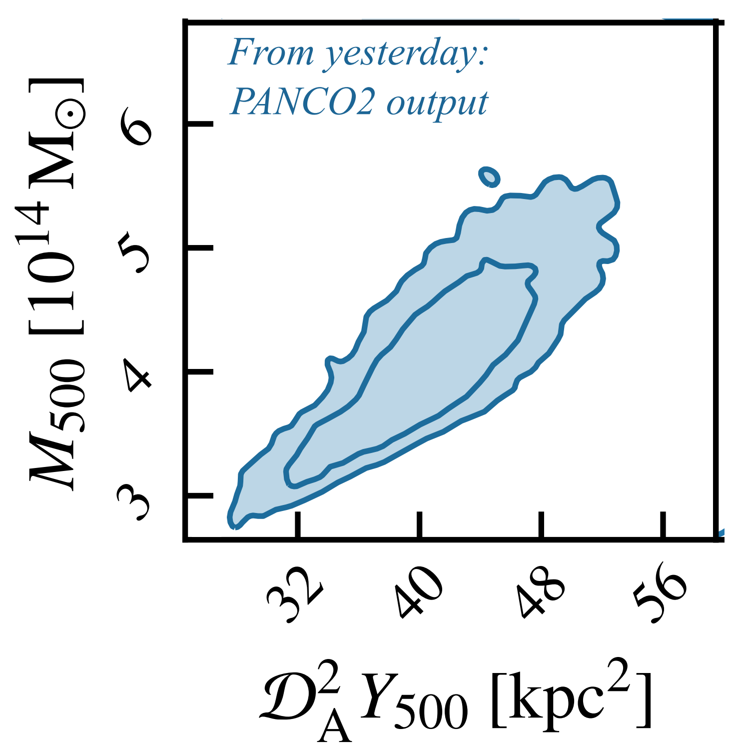
- **Step 2:** apply fiducial input SR → observable Y values
  - *Planck* results as truth:
    $\alpha_{Y|Z} = -0.19,\ \beta_{Y|Z} = 1.79,\ \sigma_{Y|Z} = 0.075$

- **Step 3:** apply the box-filling algorithm to select an LPSZ-like sample
  - Same (Y, z) bins as the real LPSZ
  - 5 clusters/box → 50 total (~real LPSZ)

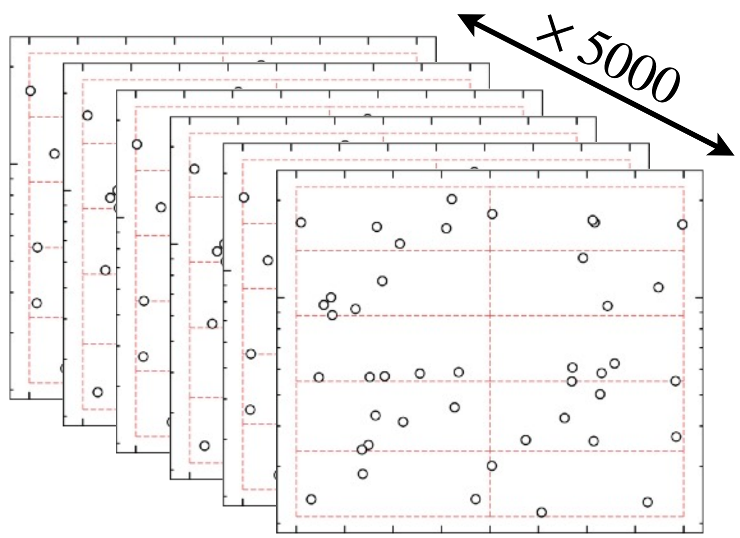
- **Step 4:** add uncertainties
  - Uncertainties on both axes and their covariance: output of individual cluster analyses (yesterday's talk)
  - Realistic values from previous cluster analyses and simulations: ~10-15% uncertainties, ~85% correlation

- Consider unbiased & unscattered mass estimators for now

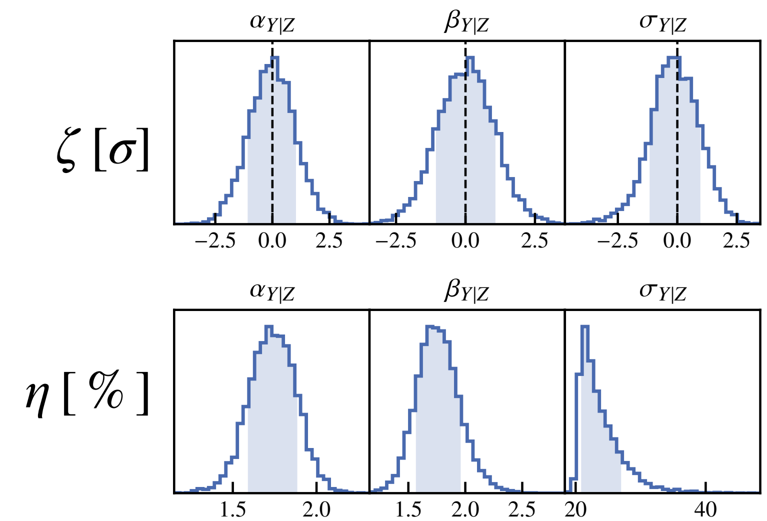




11 Accuracy & precision estimation



MCMC on each mock sample



5000 different samples generated and fitted

5000 values of ζ, η for each parameter of interest

- Repeat the procedure to generate 5000 mock samples, & fit the scaling relation on each sample
- Evaluate the bias and dispersion of the parameter estimators: for each parameter of interest ϑ with true value $\hat{\vartheta}$,

$$\text{Bias } \zeta_\vartheta\,[\sigma] \equiv \frac{\mathrm{Med}\left[\vartheta_i\right] - \hat{\vartheta}}{\sqrt{\mathrm{Var}\left[\vartheta_i\right]}} \qquad \text{Dispersion } \eta_\vartheta\,[\,\%\,] \equiv \frac{\sqrt{\mathrm{Var}\left[\vartheta_i\right]}}{|\,\hat{\vartheta}\,|}$$

(over all Markov chains samples i)

Outline

Scaling relation adjustment

Realistic mock sample generation

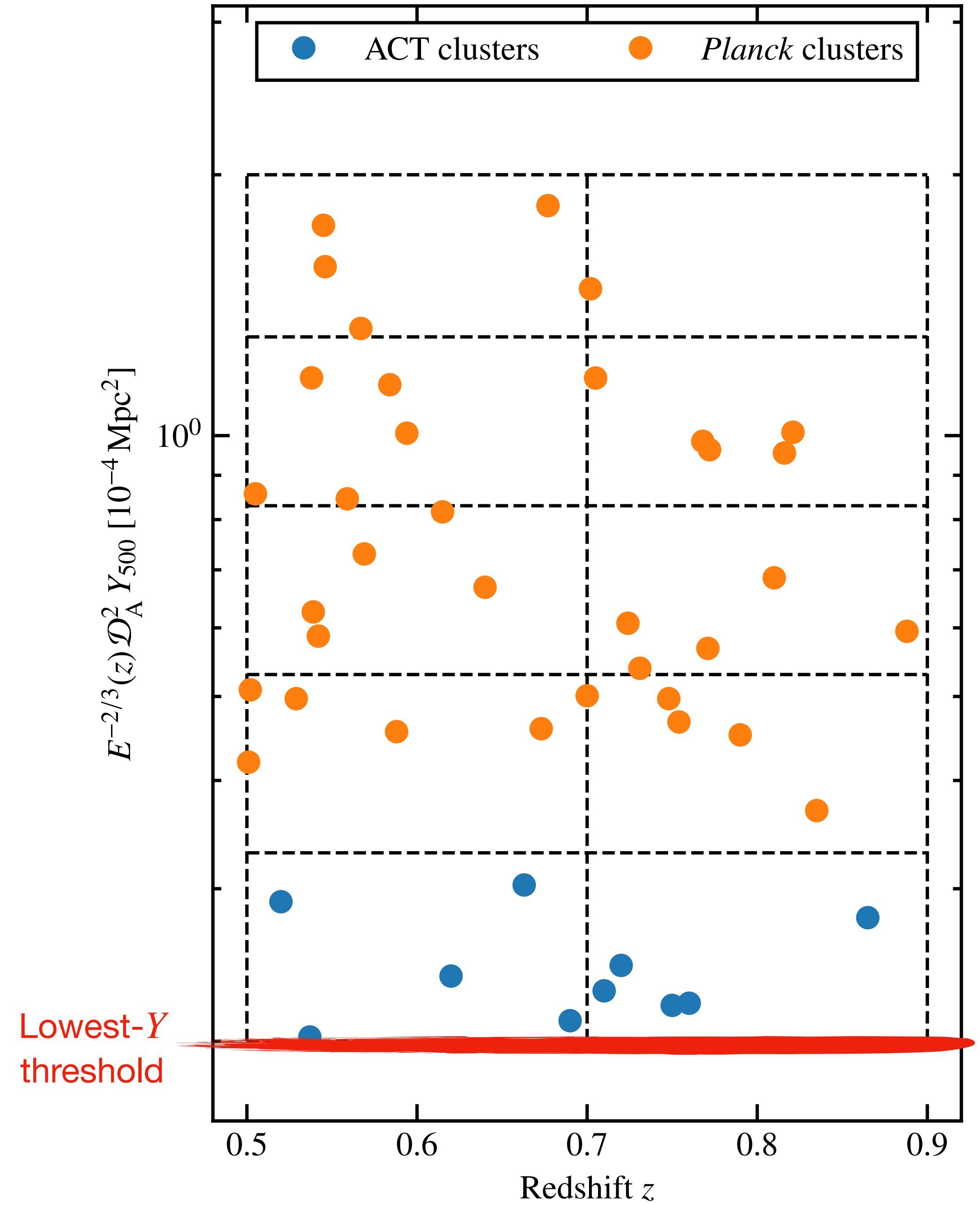
Results: biases & precision

Conclusions

LPSZ selection effects

Florian Kéruzoré
mmUniverse
June 2021

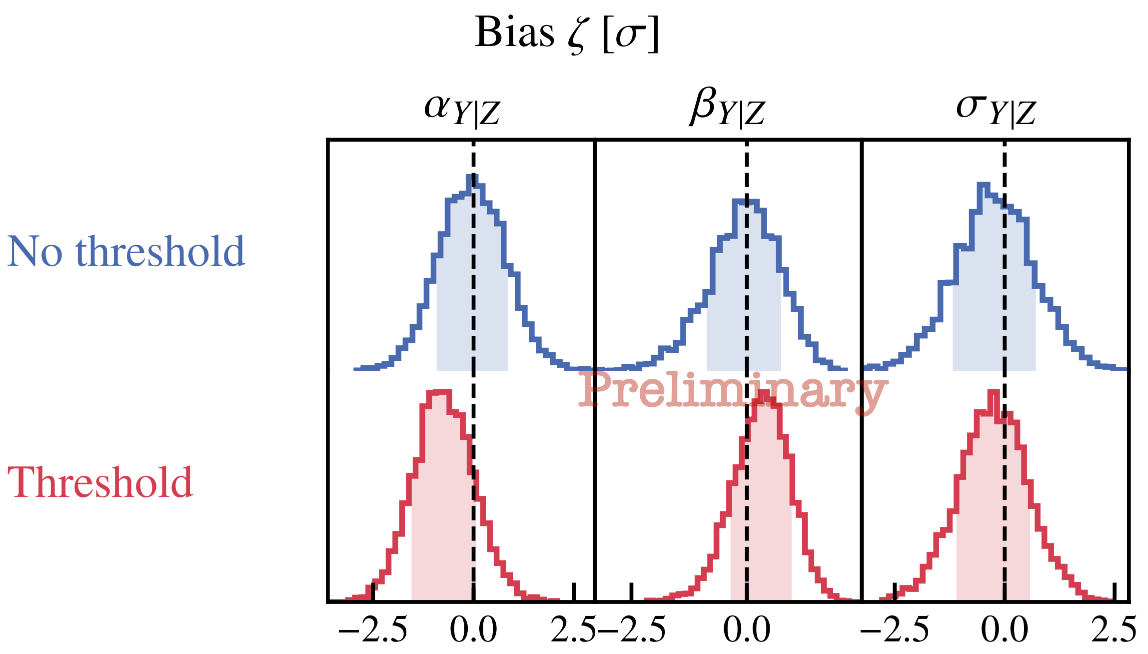
- The "box-filling" LPSZ selection is complex
  - Putting a threshold at the limit of each box is incorrect: clusters at lower values would not have been censured, just selected in a lower box
- What is the impact of the selection / how can we deal with it?
  - We could ignore the selection…
  - … Or consider a threshold at the lowest Y value
- Generate 5000 samples and fit them with both approximations



LPSZ selection effect: bias?

Florian Kéruzoré
mmUniverse
June 2021

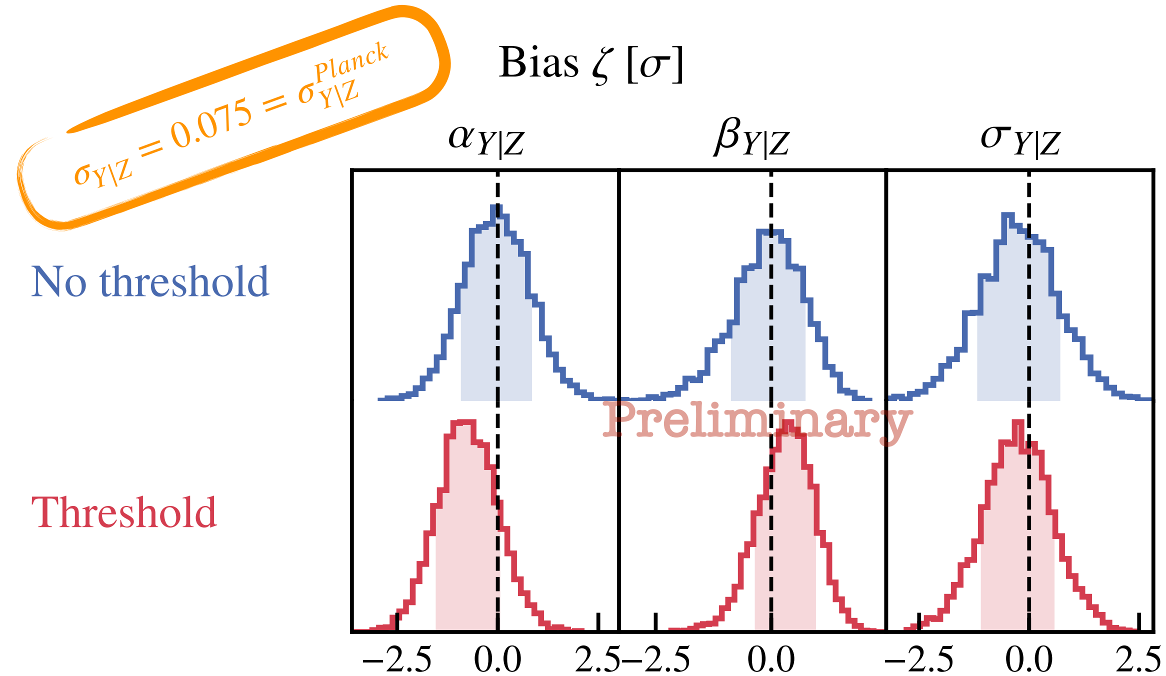
- **No significant bias** on the parameters of interest
  - Threshold in observable values: little effect
  - → No bias due to the LPSZ selection?
- Does this hold for **larger intrinsic scatter?**
  - Truth value used is low: $\sigma_{Y|Z} = 0.075$ (*Planck*)
  - Malmquist bias (MB) is due to intrinsic scatter
  - → What if we repeat with $\sigma_{Y|Z} \to 2 \times \sigma_{Y|Z}$?

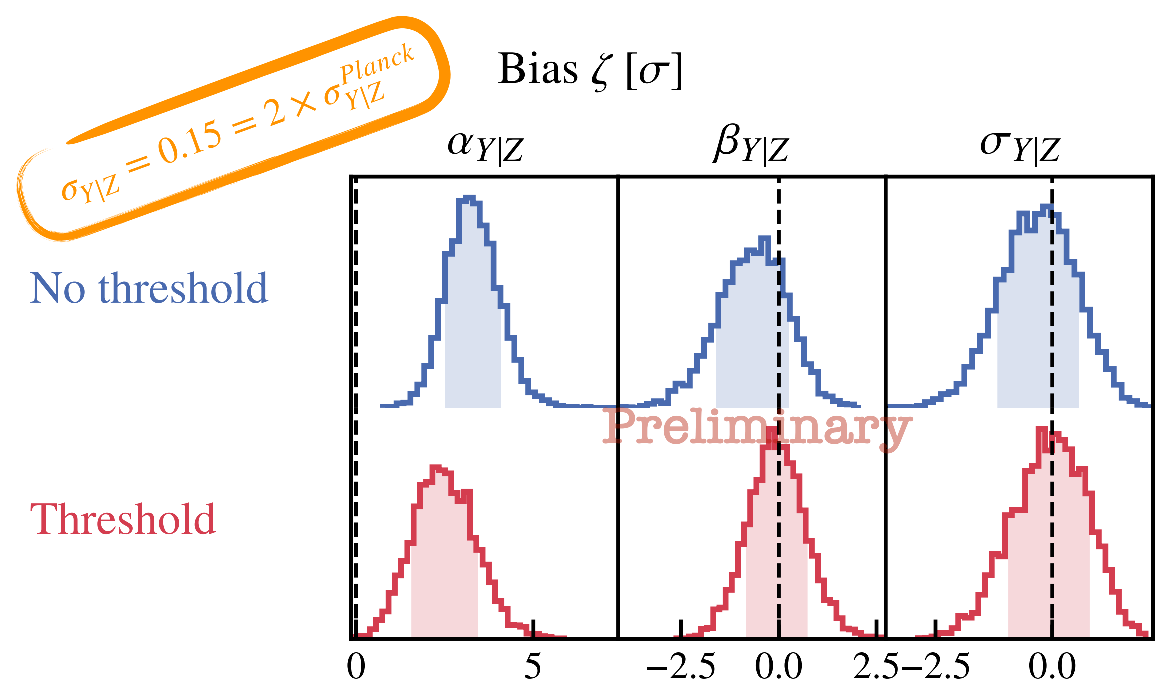


LPSZ selection effect: bias?

Florian Kéruzoré
mmUniverse
June 2021

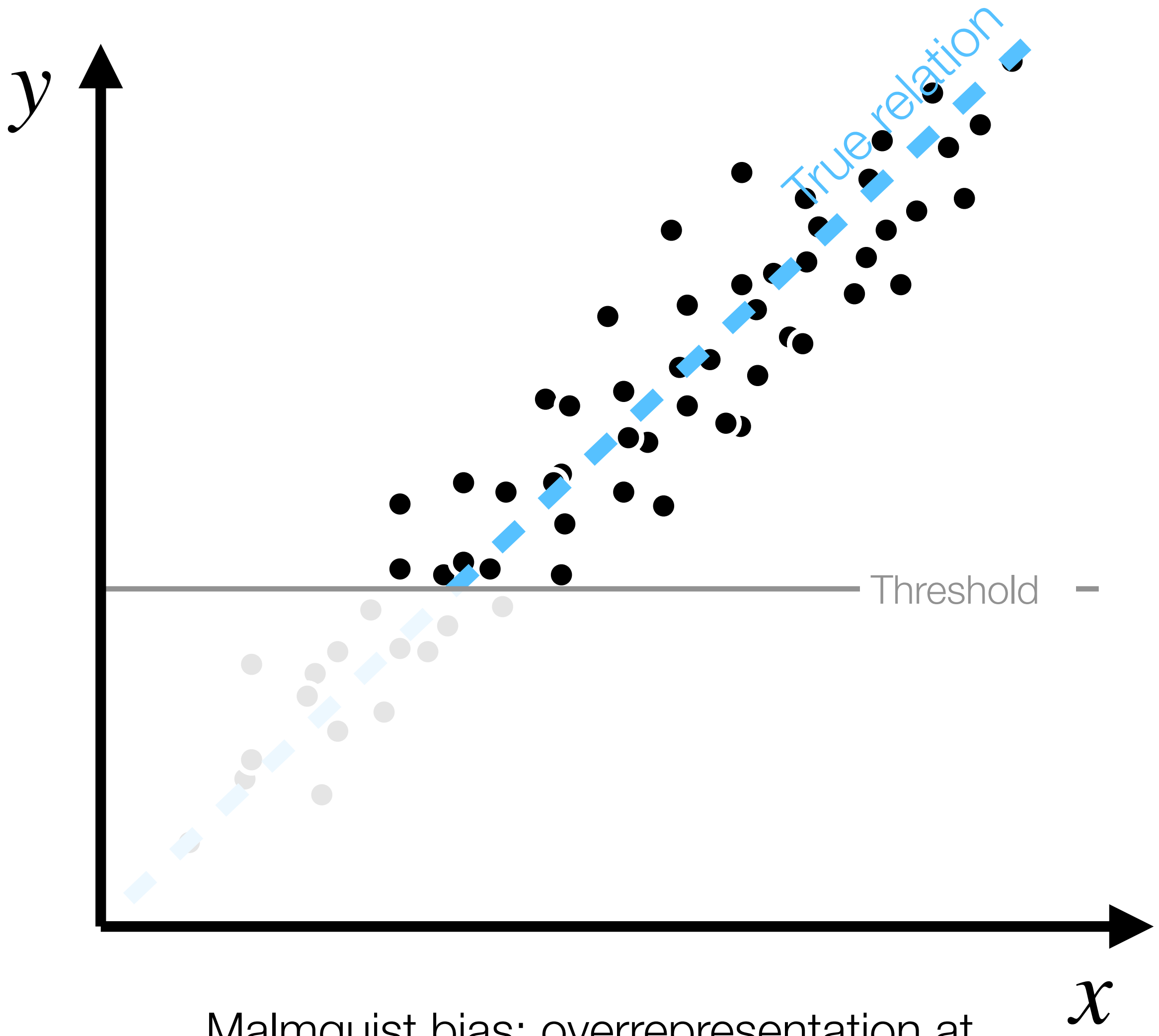
- **No significant bias** on the parameters of interest
  - Threshold in observable values: little effect
  - → No bias due to the LPSZ selection?
- Does this hold for **larger intrinsic scatter?**
  - Truth value used is low: $\sigma_{Y|Z} = 0.075$ (*Planck*)
  - Malmquist bias (MB) is due to intrinsic scatter
  - → What if we repeat with $\sigma_{Y|Z} \to 2 \times \sigma_{Y|Z}$?
- **Significant bias** on the intercept $\alpha_{Y|Z}$, with $\zeta > 2\sigma$
  - Not on the other parameters: unusual (compared to MB)
  - Considering a threshold doesn't help
- **Consequence:** if Planck underestimated intrinsic scatter, LPSZ selection creates a bias in SR measurement
  - → How can we explain this bias?
  - → Can we do something about it?





Interpretation for intercept bias

Florian Kéruzoré
mmUniverse
June 2021

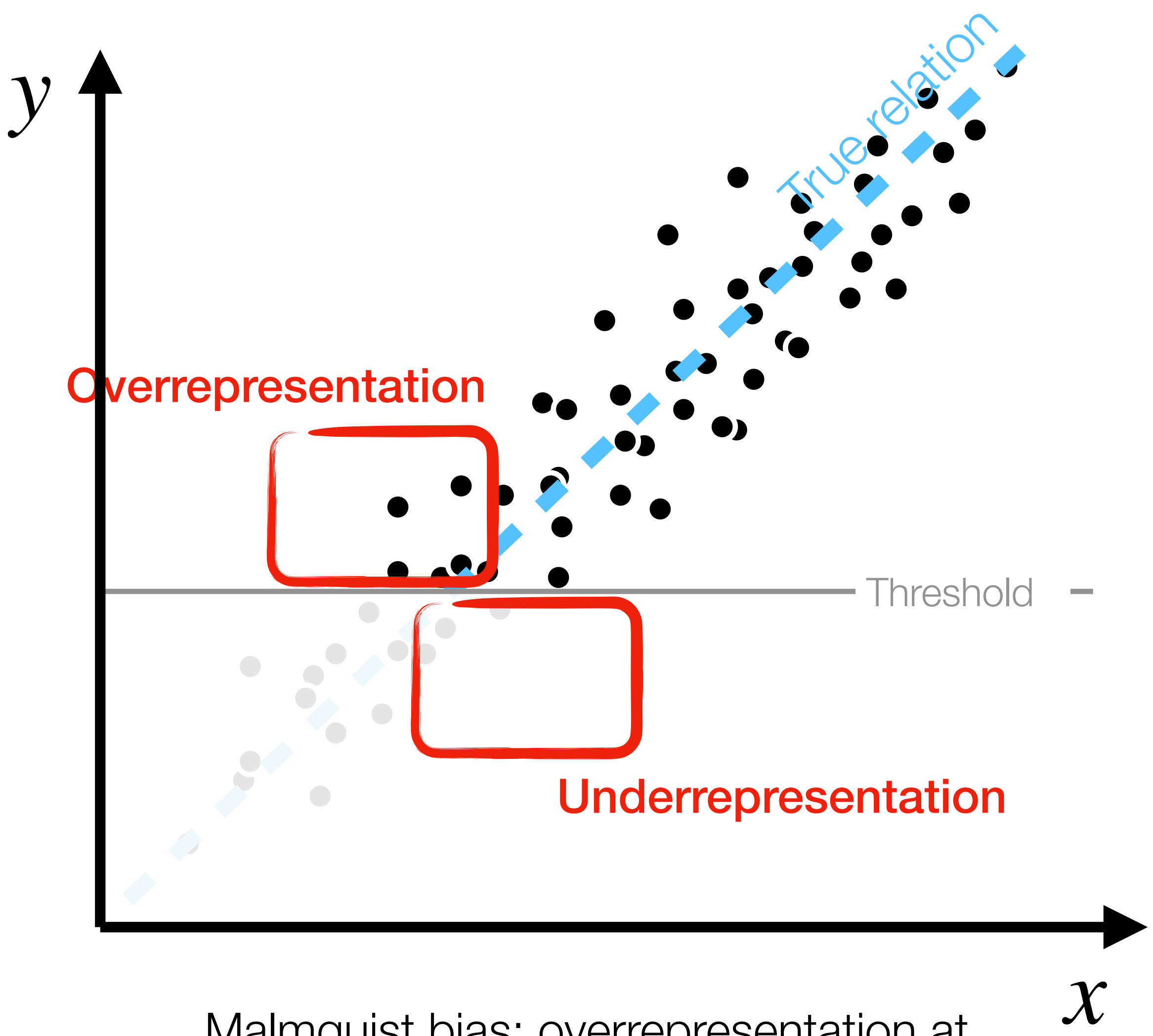


Malmquist bias: overrepresentation at detection threshold

→ Shallower relation: Biased slope

Interpretation for intercept bias

Florian Kéruzoré
mmUniverse
June 2021

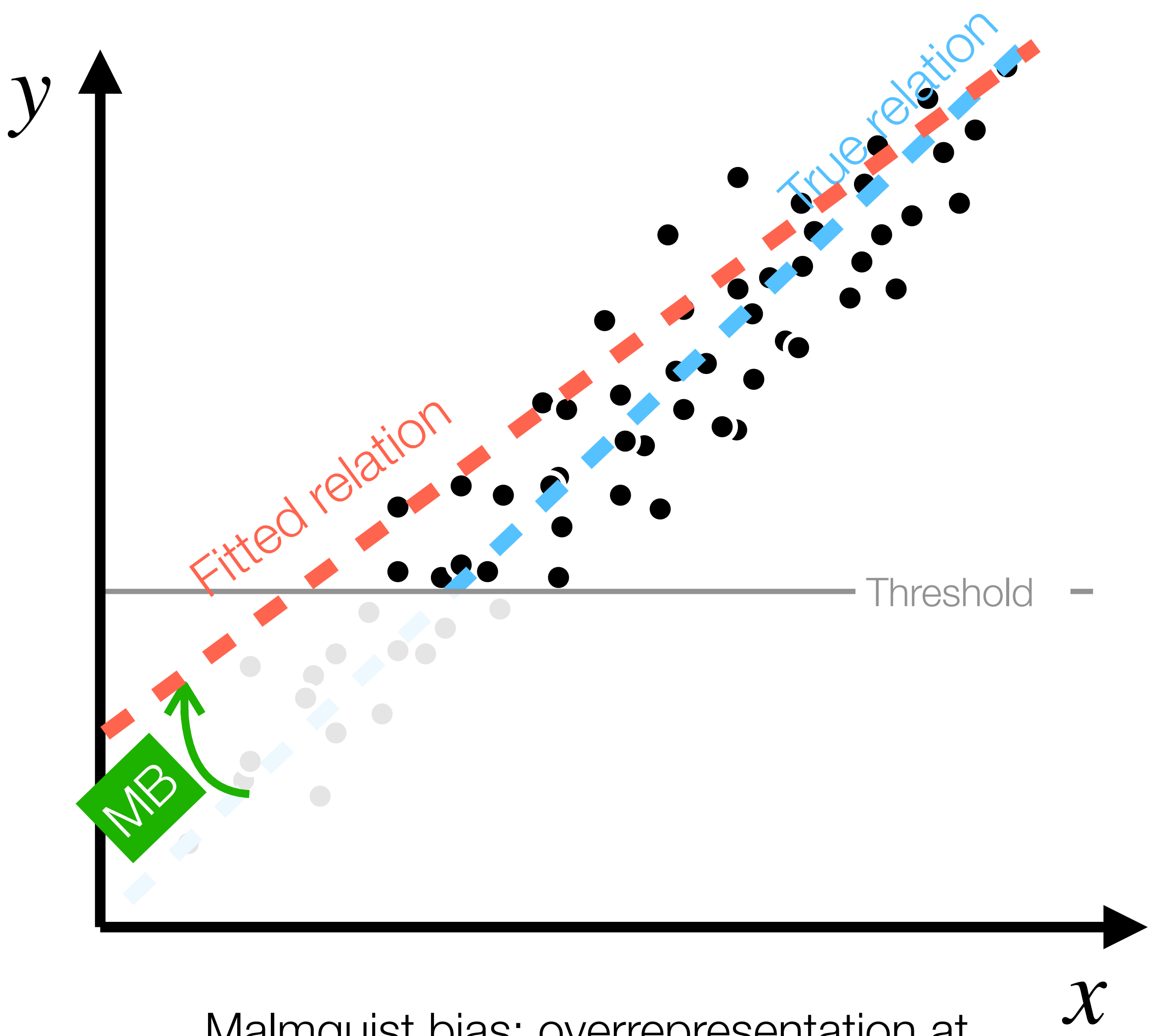


Malmquist bias: overrepresentation at detection threshold

→ Shallower relation: Biased slope

Interpretation for intercept bias

Florian Kéruzoré
mmUniverse
June 2021

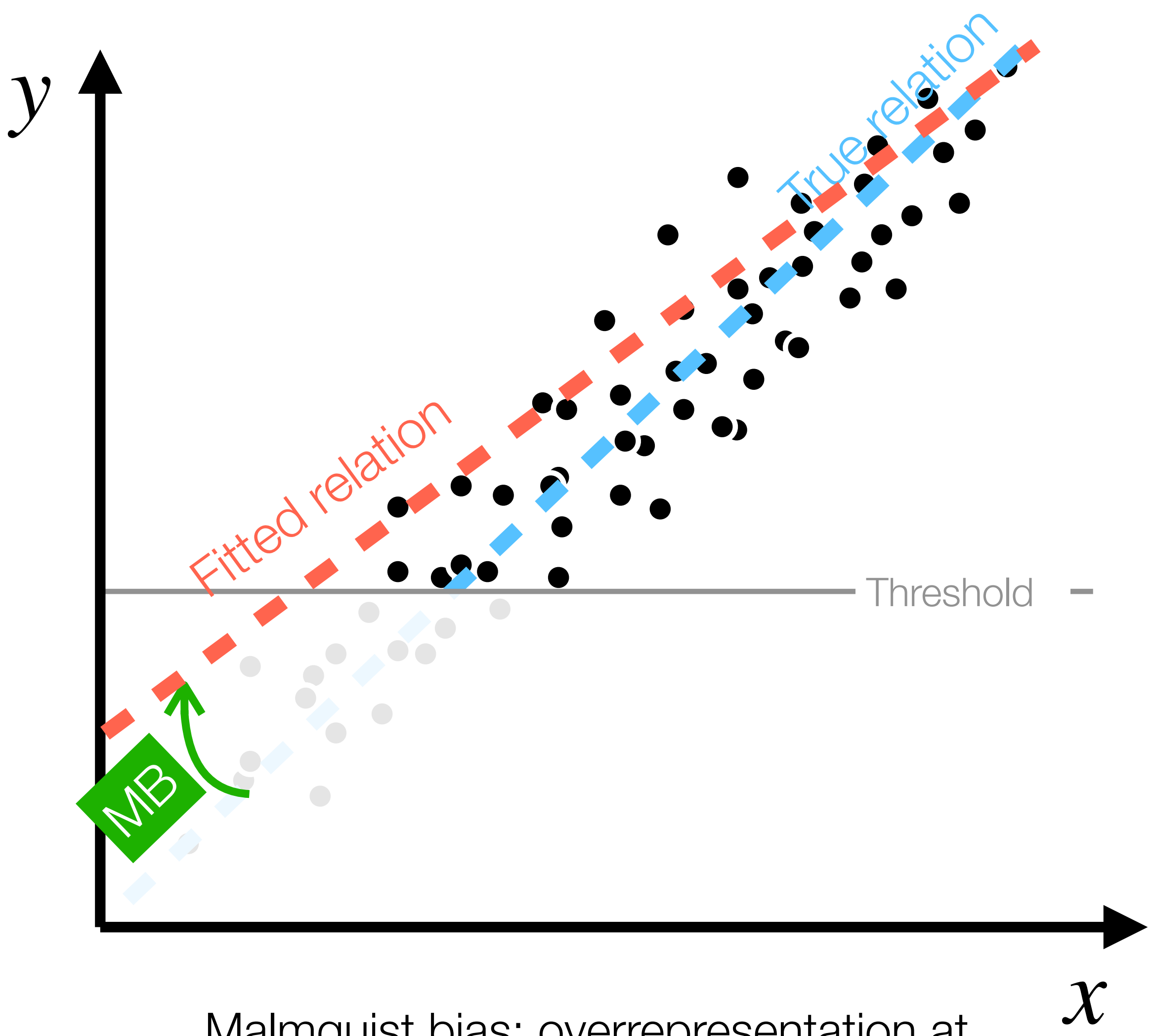


Malmquist bias: overrepresentation at detection threshold

→ Shallower relation: Biased slope

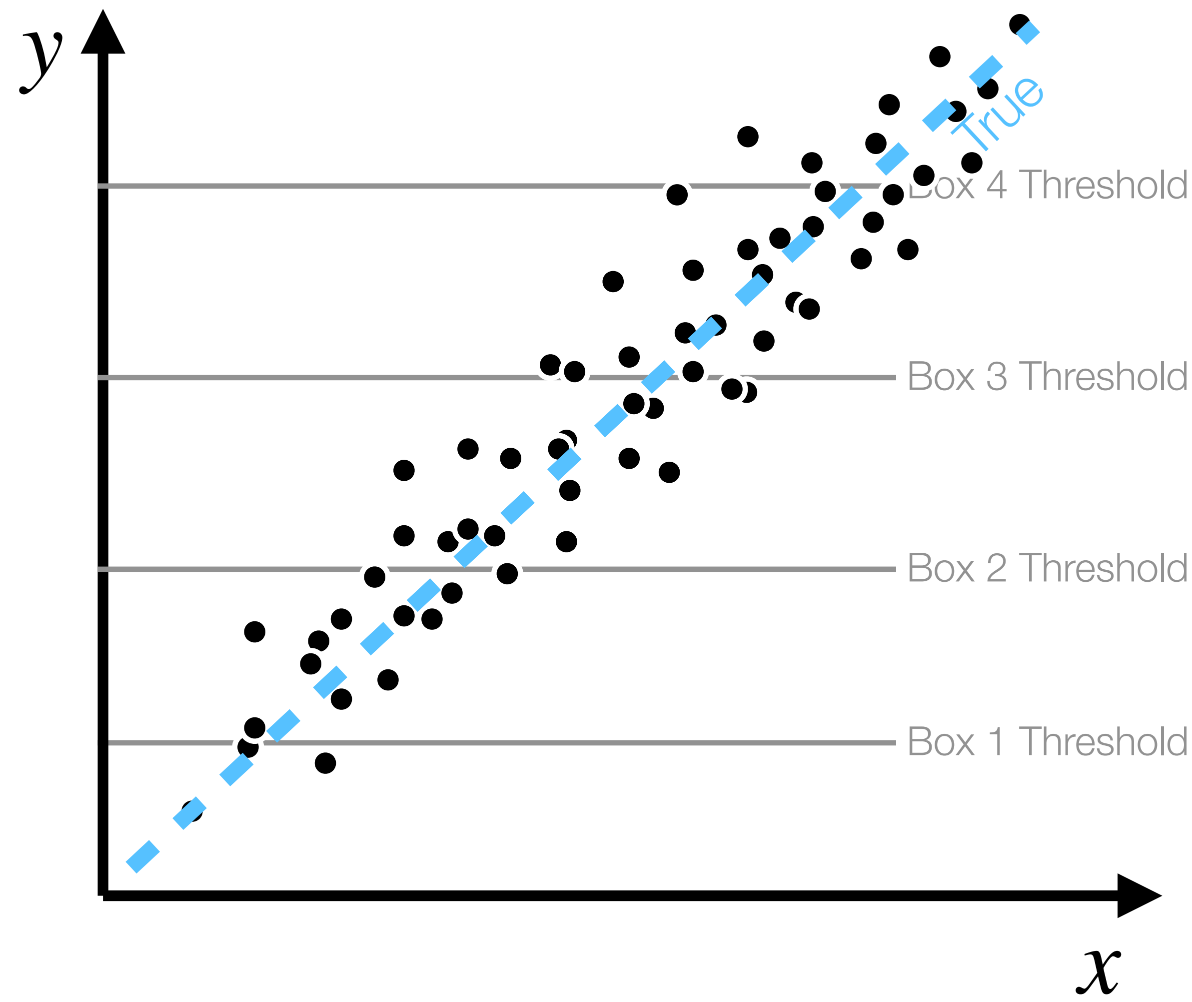
Interpretation for intercept bias

Florian Kéruzoré
mmUniverse
June 2021



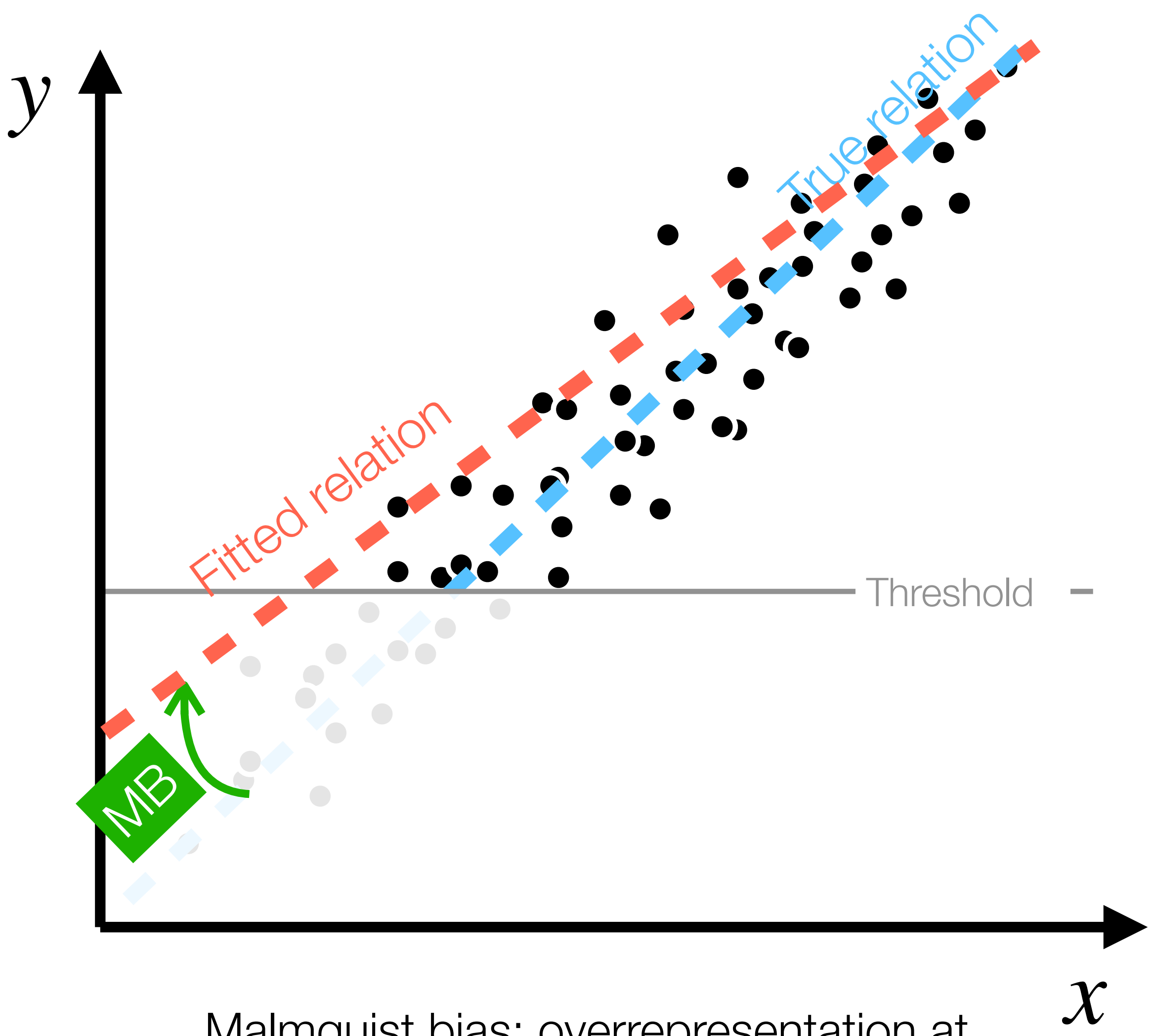
Malmquist bias: overrepresentation at detection threshold

→ Shallower relation: Biased slope



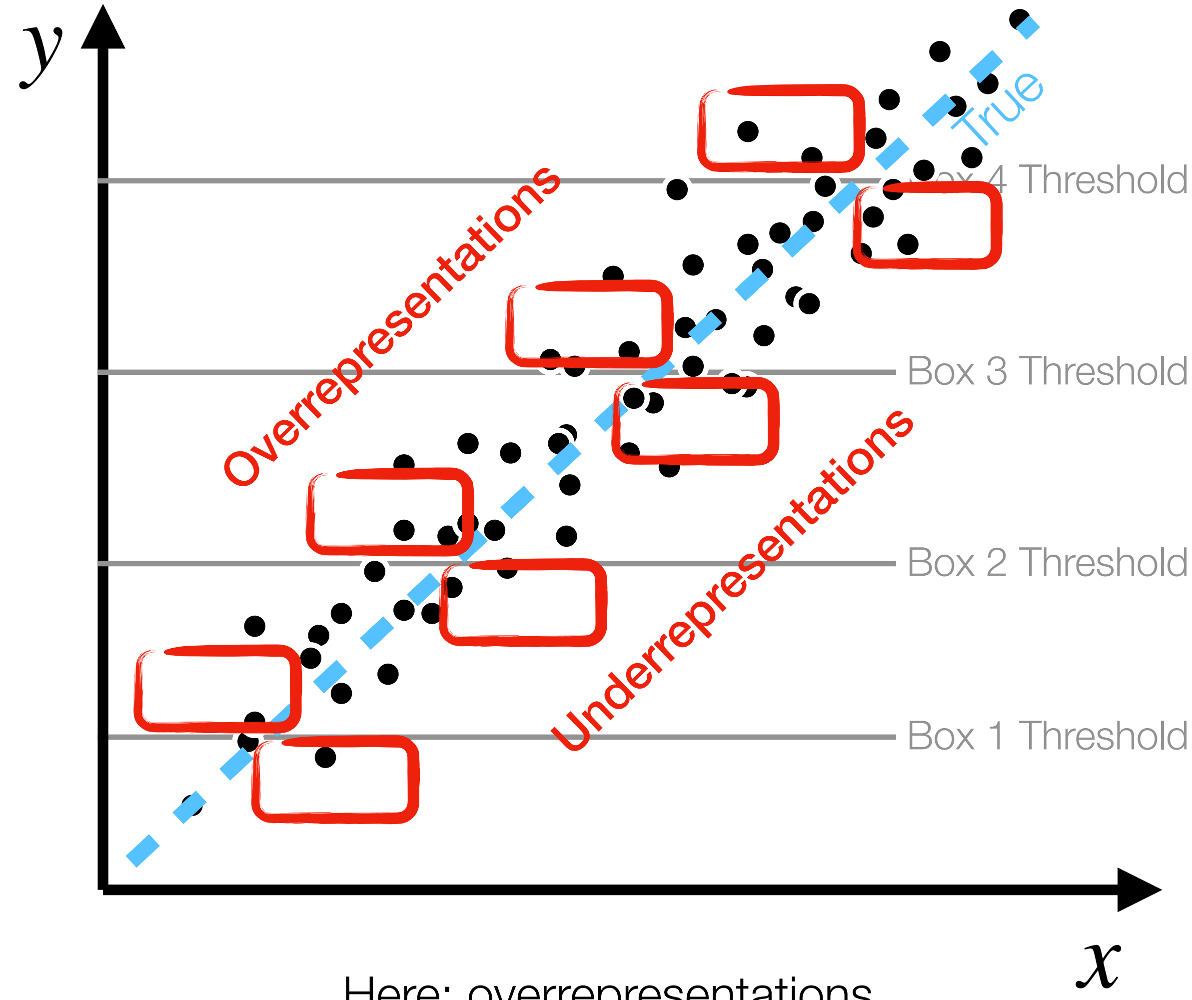
Interpretation for intercept bias

Florian Kéruzoré
mmUniverse
June 2021



Malmquist bias: overrepresentation at detection threshold

→ Shallower relation: Biased slope

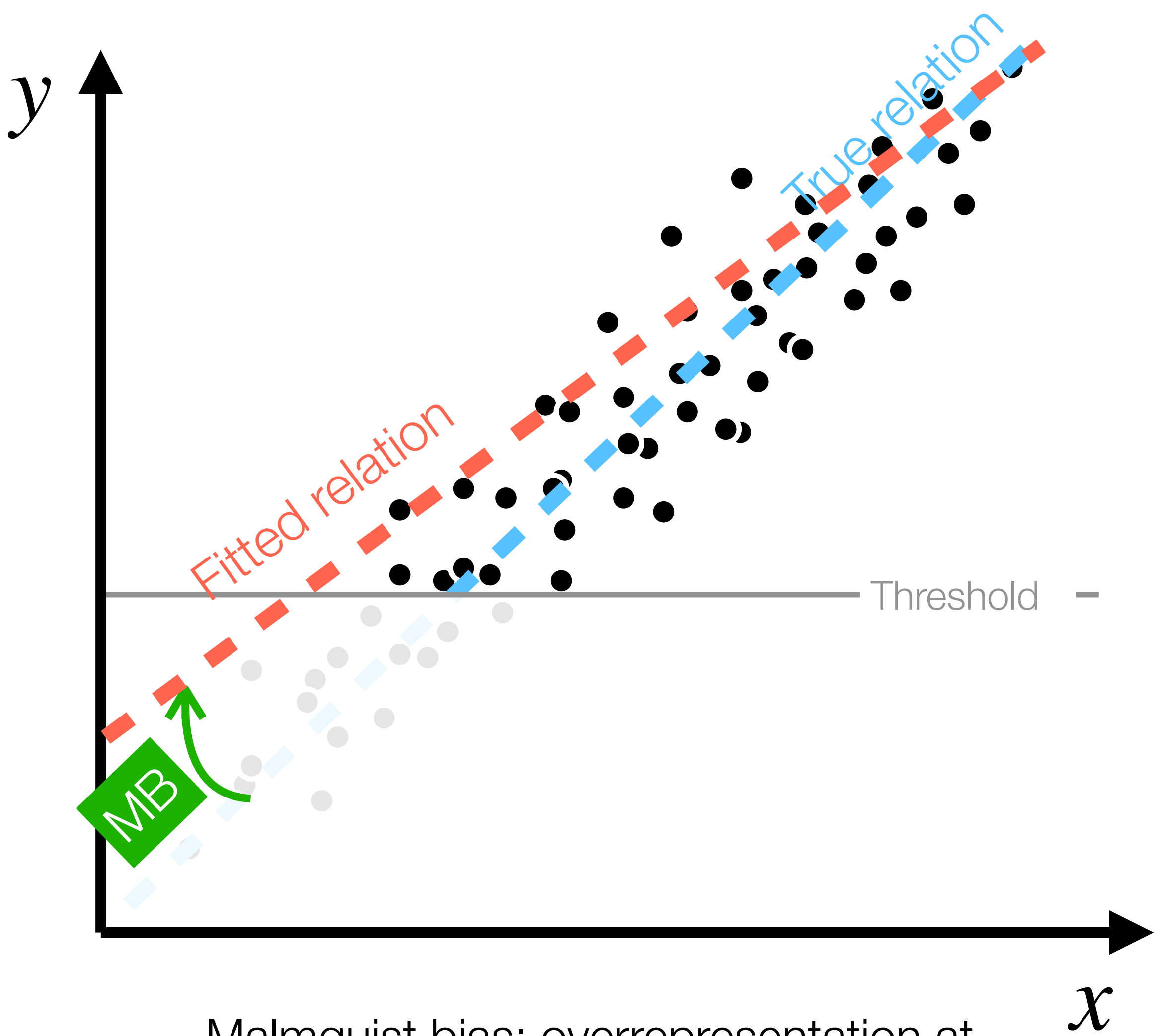


Here: overrepresentations at boxes thresholds

→ Offset relation: Biased intercept

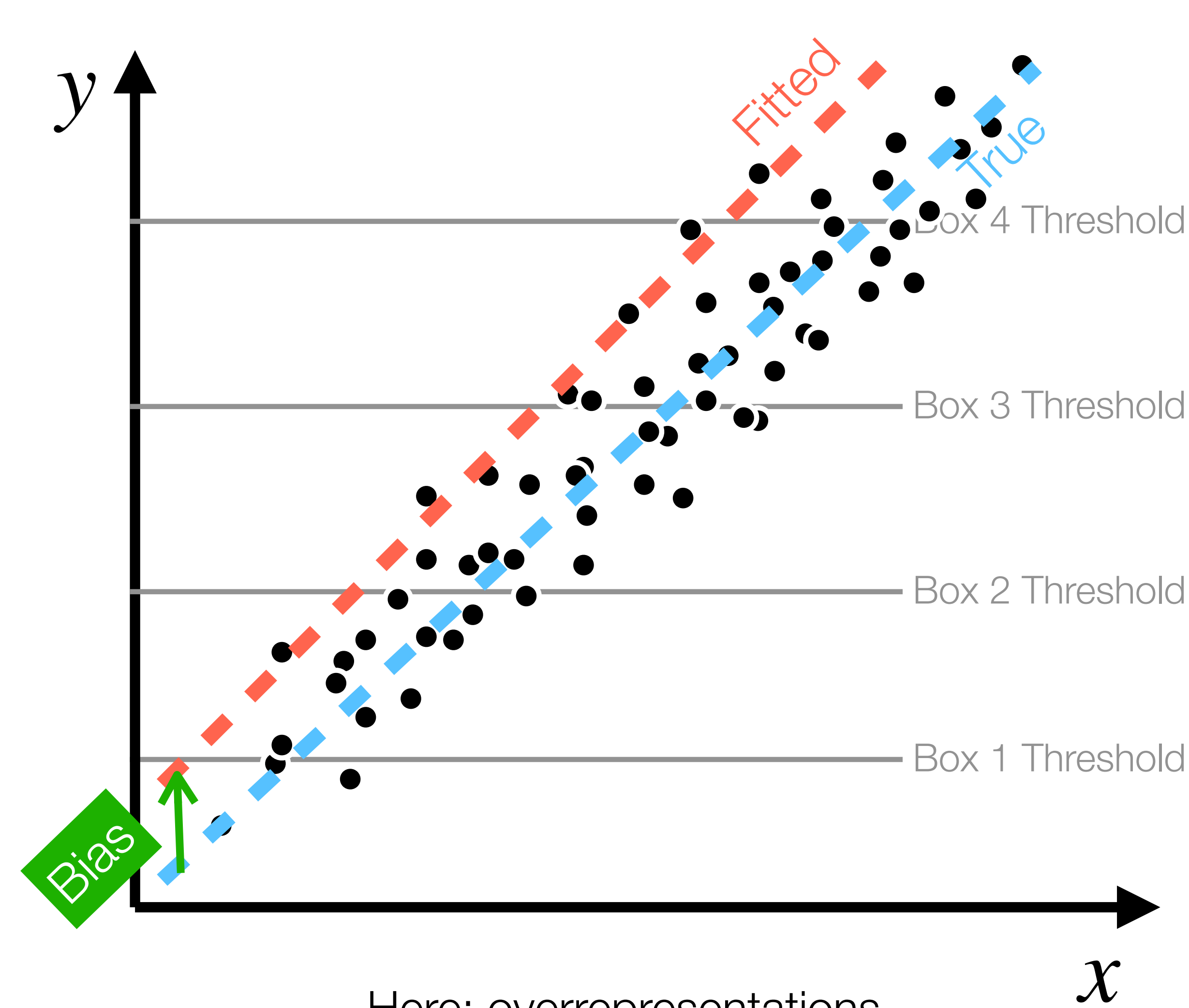
Interpretation for intercept bias

Florian Kéruzoré
mmUniverse
June 2021



Malmquist bias: overrepresentation at detection threshold

→ Shallower relation: Biased slope



Here: overrepresentations at boxes thresholds

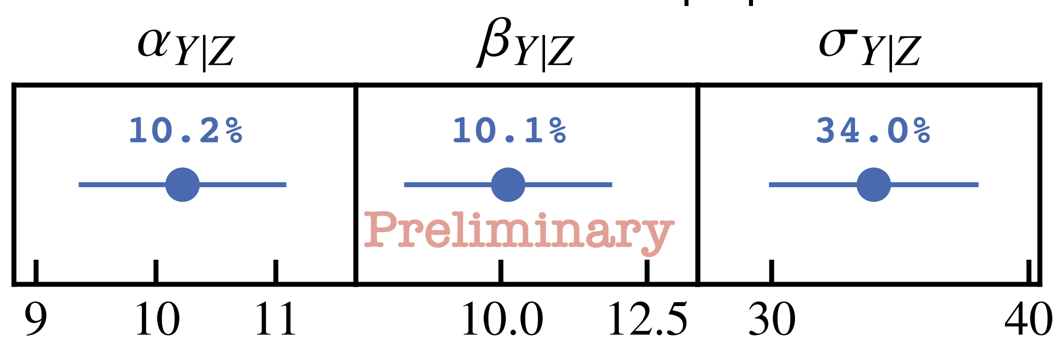
→ Offset relation: Biased intercept

- Possible solutions:
  - Study bias dependence with σ on simulations, and correct ad-hoc
  - Measure α independently and fix it in the analysis
  - If the scatter is low (as measured by Planck), bias is negligible

16 Parameter precision

Florian Kéruzoré
mmUniverse
June 2021

$$\text{Dispersion } \eta\ [\%] = \frac{\sqrt{\mathrm{Var}\left[\vartheta_i\right]}}{|\hat{\vartheta}|}$$



- For small intrinsic scatter → negligible selection bias
- Relatives uncertainties on parameters η:
  - ~10% on average on α & β
  - ~30% on σ

Outline

Scaling relation adjustment

Realistic mock sample generation

Results: biases & precision

Conclusions

Summary & conclusions

- $Y_{500} - M_{500}$ scaling relation = NIKA2 LPSZ goal
- Constraining power evaluated on mock datasets
  - Generated with a realistic procedure
  - Fitted with a Bayesian hierarchical model using the `LIRA` library
- Results: bias and dispersion of the parameter estimators
  - LPSZ selection creates bias in the SR intercept, not on other parameters
  - Negligible for low intrinsic scatter (as measured by *Planck*)
  - Dispersion around 10% for scaling relation parameters
- Main assumptions/caveats:
  - mass bias/dispersion not accounted for yet
  - input survey selection not accounted for yet
- Forecasting and decision help for future sample studies using NIKA2