Partial Galaxy Clustering: An Estimator Incorporating Probabilistic Distance Measurements

Humna Awan

Advisor: Eric Gawiser

Rutgers University, Dept. of Physics & Astronomy

April 20, 2018
SCLSS, Oxford

De-Projection

Consider how the correlations in the contaminated subsamples relate to the true ones:

$$\begin{bmatrix} w_{LL}^{obs} \\ w_{LO}^{obs} \\ w_{OO}^{obs} \end{bmatrix} = \begin{bmatrix} f_{LL}^2 & 2f_{LL}f_{LO} & f_{LO}^2 \\ f_{LL}f_{OL} & f_{LL}f_{OO} + f_{OL}f_{LO} & f_{OO}f_{LO} \\ f_{OL}^2 & 2f_{OO}f_{OL} & f_{OO}^2 \end{bmatrix} \begin{bmatrix} w_{LL}^{true} \\ w_{LO}^{true} \\ w_{OO}^{true} \end{bmatrix}$$

<u>**Assumes**</u> the classification probabilities can be represented by their sample averages.

De-Projection

Consider how the correlations in the contaminated subsamples relate to the true ones:

$$\begin{bmatrix} w_{LL}^{obs} \\ w_{LO}^{obs} \\ w_{OO}^{obs} \end{bmatrix} = \begin{bmatrix} f_{LL}^2 & 2 f_{LL} f_{LO} & f_{LO}^2 \\ f_{LL} f_{OL} & f_{LL} f_{OO} + f_{OL} f_{LO} & f_{OO} f_{LO} \\ f_{OL}^2 & 2 f_{OO} f_{OL} & f_{OO}^2 \end{bmatrix} \begin{bmatrix} w_{LL}^{true} \\ w_{LO}^{true} \\ w_{OO}^{true} \end{bmatrix}$$

Assumes the classification probabilities can be represented by their sample averages.

=> De-projected LS estimators for the auto/cross-correlations:

$$\begin{bmatrix} \widehat{w}_{LL} \\ \widehat{w}_{LO} \\ \widehat{w}_{OO} \end{bmatrix} = \begin{bmatrix} f_{LL}^2 & 2 f_{LL} f_{LO} & f_{LO}^2 \\ f_{LL} f_{OL} & f_{LL} f_{OO} + f_{OL} f_{LO} & f_{OO} f_{LO} \\ f_{OL}^2 & 2 f_{OO} f_{OL} & f_{OO}^2 \end{bmatrix}^{-1} \begin{bmatrix} w_{LL}^{obs} \\ w_{LO}^{obs} \\ w_{OO}^{obs} \end{bmatrix}$$

Possible improvement to assumptions about contamination?

Estimators that incorporate uncertainty in galaxy radial positions

Probability-Weighted Estimator

Marked correlations: extract features in correlations.

Weigh each galaxy by its classification probability!

$\Rightarrow$ Consider *all* galaxies, without divisions into subsamples.

$\Rightarrow$ Probability-weighted estimator

$$\widetilde{w}^{obs}_{AB}(\theta_k) = \frac{(\widetilde{DD})_{AB}(\theta_k) - (\widetilde{DR})_{A}(\theta_k) - (\widetilde{DR})_{B}(\theta_k) + RR(\theta_k)}{RR(\theta_k)}$$

where

$$(\widetilde{DD})_{AB}(\theta_k) \sim \sum_{i}^{N_{tot}} \sum_{j \neq i}^{N_{tot}} w_i^A w_j^B \Theta(\theta_{ij} - \theta_{\min,k})[1 - \Theta(\theta_{ij} - \theta_{\max,k})]$$

$$(\widetilde{DR})_{A}(\theta_k) \sim \sum_{i}^{N_{tot}} \sum_{j}^{N_R} w_i^A \Theta(\theta_{ij} - \theta_{\min,k})[1 - \Theta(\theta_{ij} - \theta_{\max,k})]$$

Probability-Weighted Estimator: De-Biasing

$$\widetilde{w}^{obs}_{AB}(\theta_k) = \frac{(\widetilde{DD})_{AB}(\theta_k) - (\widetilde{DR})_A(\theta_k) - (\widetilde{DR})_B(\theta_k) + RR(\theta_k)}{RR(\theta_k)}.$$

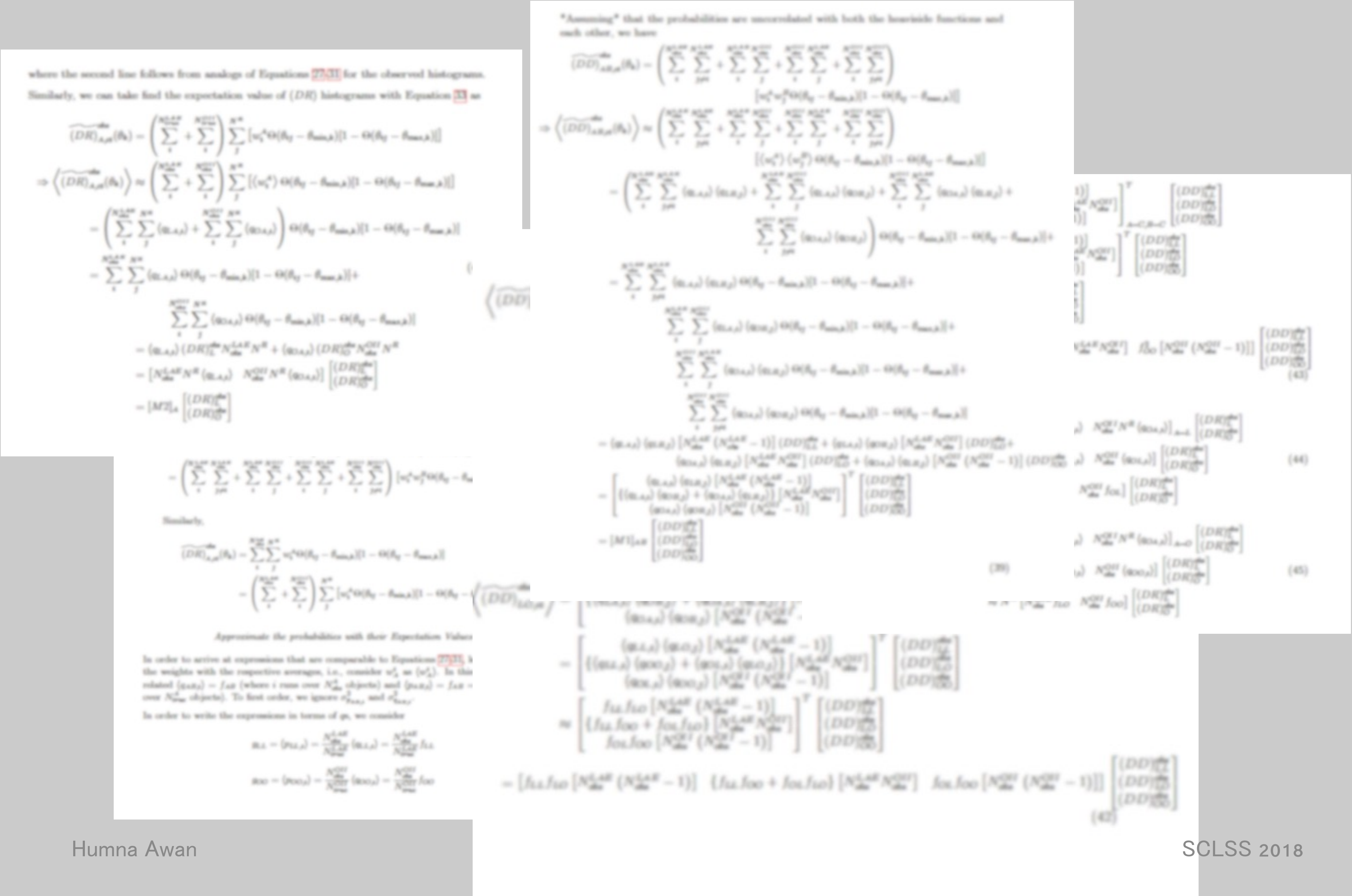
$\widetilde{w}^{obs}_{AB}$ is biased: need to de-bias to get $\widehat{w}$

We have

$$\begin{bmatrix} \widehat{w}_{LL} \\ \widehat{w}_{LO} \\ \widehat{w}_{OO} \end{bmatrix} = \frac{1}{RR} \begin{bmatrix} 1 & 0 & 0 & -2 & 0 & 1 \\ 0 & 1 & 0 & -1 & -1 & 1 \\ 0 & 0 & 1 & 0 & -2 & 1 \end{bmatrix} \begin{bmatrix} (\widehat{DD})_{LL} \\ (\widehat{DD})_{LO} \\ (\widehat{DD})_{OO} \\ (\widehat{DR})_{L} \\ (\widehat{DR})_{O} \\ RR \end{bmatrix}$$

$\Rightarrow$ Can de-bias individual histograms, $(\widetilde{DD})_{AB}$, $(\widetilde{DR})_A$

Probability-Weighted Estimator: De-Biasing



Probability-Weighted Estimator: De-Biasing

After all the algebra and some simplifications, we have

$$\begin{bmatrix} \widehat{(DD)}_{LL} \\ \widehat{(DD)}_{LO} \\ \widehat{(DD)}_{OO} \\ \widehat{(DR)}_{L} \\ \widehat{(DR)}_{O} \\ RR \end{bmatrix} \equiv \begin{bmatrix} \langle (DD)^{true}_{LL} \rangle \\ \langle (DD)^{true}_{LO} \rangle \\ \langle (DD)^{true}_{OO} \rangle \\ \langle (DR)^{true}_{L} \rangle \\ \langle (DR)^{true}_{O} \rangle \\ RR \end{bmatrix} = \{[M][C]\}^{-1} \begin{bmatrix} \widetilde{(DD)}^{obs}_{LL} \\ \widetilde{(DD)}^{obs}_{LO} \\ \widetilde{(DD)}^{obs}_{OO} \\ \widetilde{(DR)}^{obs}_{L} \\ \widetilde{(DR)}^{obs}_{O} \\ RR \end{bmatrix}$$

with [M], [C] are calculable given the weights.

Test

We apply the estimators to a HETDEX mock catalog**

- **2-sample case:** either one is a contaminant w.r.t the other.
- Can construct a **probabilistic classifier** assigning each observed galaxy of type A a probability of being type B: q_{AB}
- Use the probabilities in the estimators!

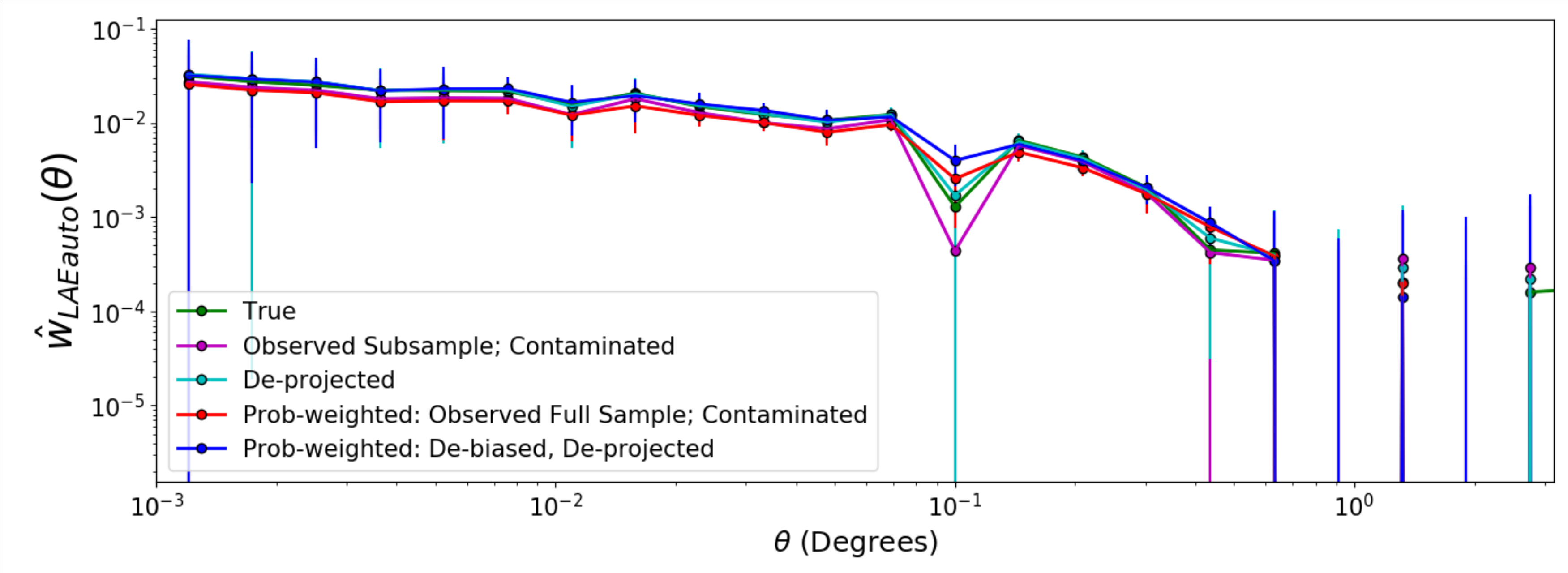
  Renders each galaxy's existence in a sample a probabilistic existence in each distance bin.

- Example realization: 719,881 true LAEs and 465,104 true [OII] emitters
- Implement 10% LAE sample contamination; 6% incompleteness to create observed catalogs.
- Well-behaved, unbiased classification probability distributions.

- Jackknife to get the variance (while work in progress for analytical expressions)

*Thanks to Chi-Ting Chiang.

Results: LAE auto-correlation

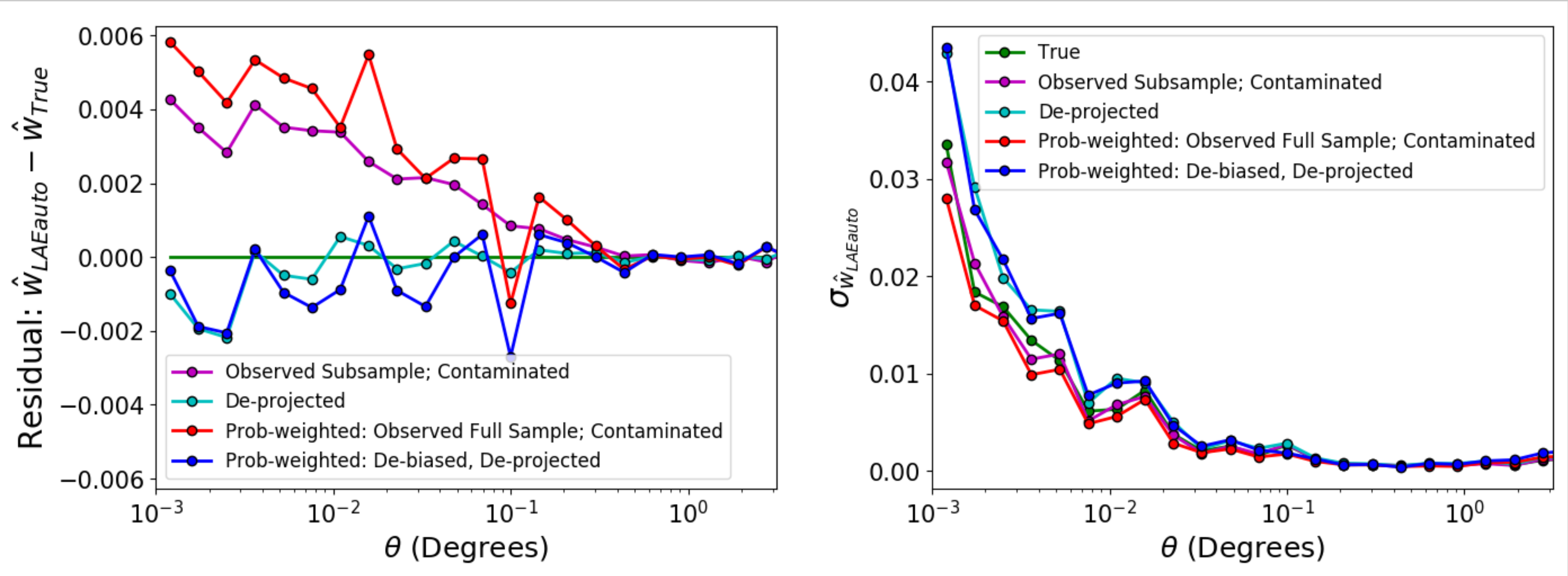


Awan & Gawiser, in prep

Weights for each galaxy= classification probability

Jackknife errors

Results: LAE auto-correlation



Awan & Gawiser, in prep

Weights for each galaxy= classification probability

New estimator gives unbiased result => de-biasing is working.
Variance is comparable with simplest weights.

Summary

- Improved galaxy clustering estimators:
  - Needed to account for measurement uncertainties directly.
  - Photo-z surveys, e.g. LSST: ~9-contaminant case. 2D.
  - Emission-line surveys, e.g. HETDEX: 1-contaminant case. 3D.
- Discussed here: probability-weighted estimator
  - Uses probabilistic distance measurements.
  - **Have the infrastructure to test different weights.**

Current Work

- Optimize weights to minimize/reduce variance.
- Apply the estimators to a photo-z catalog: 2D applicable.
  - De-biasing+variance for general classification prob. distributions.
  - Extend 2-sample methods to 3-sample (then generalizable?).

Future

- Estimators for 3D correlations.

Thanks to RDI[2] Fellowship for Excellence in Computation and Data Science 2017-2018

Galaxy Correlation Functions

2pt galaxy autocorrelation function *$w(\theta)$* (angular= 2D)

- A common statistic to study galaxy clustering
- Measures excess probability of finding a galaxy at an angular distance θ from another galaxy in comparison with a random distribution: $dP = n[1 + w(\theta)]d\Omega$

Galaxy Clustering: Traditional Estimator

(2D) 2pt galaxy autocorrelation function *w(θ)*

- Landy-Szalay estimator:

$$w_{auto}(\theta) = \frac{(D-R)(D-R)(\theta)}{RR(\theta)} = \frac{DD(\theta) - 2DR(\theta) + RR(\theta)}{RR(\theta)}$$

DD, DR, RR are histograms.

Explicitly, e.g. ,

$$DD(\theta_k) = \frac{\sum_i^N \sum_{j>i}^N \Theta(\theta_{ij} - \theta_{\min,k})[1 - \Theta(\theta_{ij} - \theta_{\max,k})]}{\sum_i^N \sum_{j>i}^N}$$

where $\Theta(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases}$ is the Heaviside step function.

Galaxy Clustering: Traditional Estimator

(2D) 2pt galaxy autocorrelation function *w(θ)*

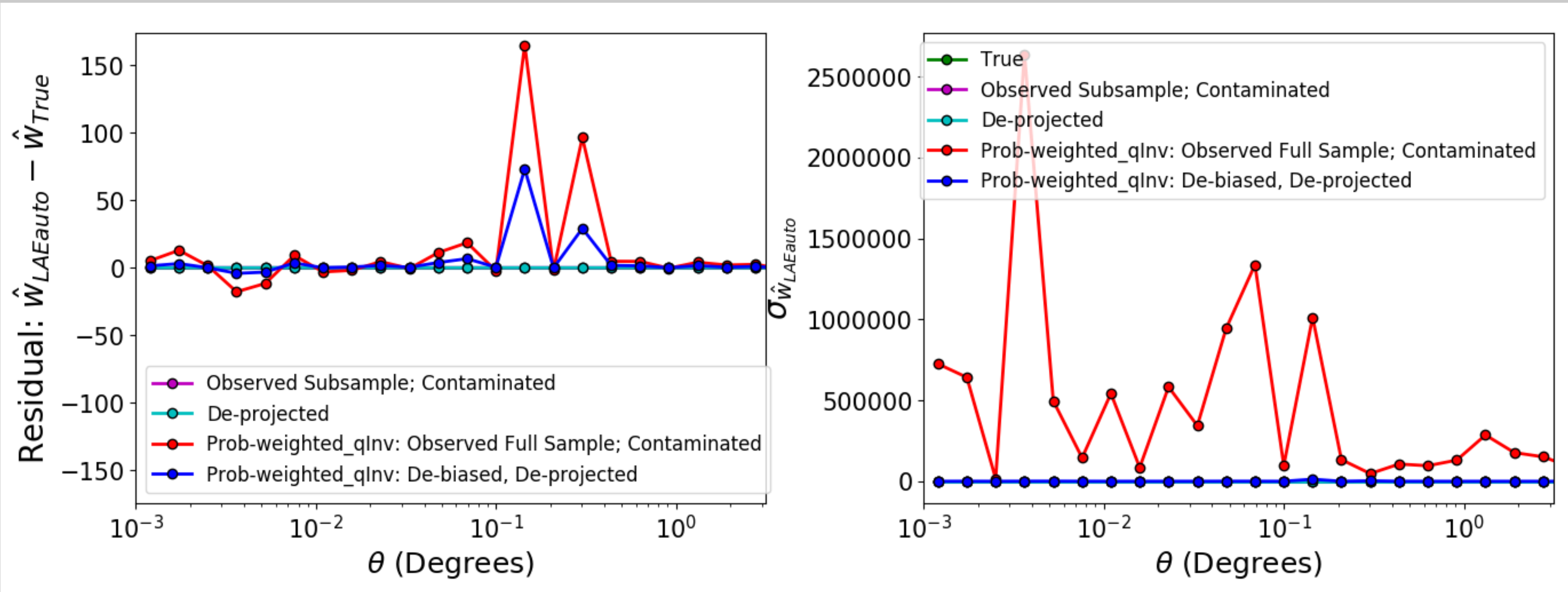
- Landy-Szalay estimator:

$$w_{auto}(\theta) = \frac{(D-R)(D-R)(\theta)}{RR(\theta)} = \frac{DD(\theta) - 2DR(\theta) + RR(\theta)}{RR(\theta)}$$

Unbiased estimator but requires a “clean” sample
$\Rightarrow$ Need to make assumptions about the contamination in the sample -- limits utilizing all the available information.

Why is it a problem?

Results: LAE auto-correlation



Awan & Gawiser, in prep

Sanity check:

Weights for each galaxy= 1/(classification probability)

Expect things to not work, and they don’t.