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:

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

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

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

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

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 <u>classification probability</u>! \Rightarrow Consider *all* galaxies, without divisions into subsamples. \Rightarrow Probability-weighted estimator

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

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

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

 \widetilde{w}_{AB}^{obs} 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

Assuming that the probabilities are uncorrelated with both the hearinide functions and rach other, we have

 $= \begin{bmatrix} f_{LL}f_{LO} \left[N_{dis}^{LAE} \left(N_{dis}^{LAE} - 1 \right) \right] \quad \left\{ f_{LL}f_{DO} + f_{OL}f_{LO} \right\} \begin{bmatrix} N_{dis}^{LAE} N_{dis}^{OUI} \end{bmatrix} \quad f_{OL}f_{DO} \begin{bmatrix} N_{dis}^{OUI} \left(N_{dis}^{OUI} - 1 \right) \right] \quad \left\{ f_{LL}f_{DO} + f_{OL}f_{LO} \right\} \begin{bmatrix} N_{dis}^{LAE} N_{dis}^{OUI} \end{bmatrix} \quad f_{OL}f_{DO} \begin{bmatrix} N_{dis}^{OUI} \left(N_{dis}^{OUI} - 1 \right) \right] \quad \left\{ f_{LL}f_{DO} + f_{OL}f_{LO} \right\} \begin{bmatrix} N_{dis}^{LAE} N_{dis}^{OUI} \end{bmatrix} \quad f_{OL}f_{DO} \begin{bmatrix} N_{dis}^{OUI} \left(N_{dis}^{OUI} - 1 \right) \right] \quad \left\{ f_{LL}f_{DO} + f_{OL}f_{LO} \right\} \begin{bmatrix} N_{dis}^{LAE} N_{dis}^{OUI} \end{bmatrix} \quad f_{OL}f_{DO} \begin{bmatrix} N_{dis}^{OUI} \left(N_{dis}^{OUI} - 1 \right) \right] \quad \left\{ f_{LL}f_{DO} + f_{DO}f_{LO} \right\} \begin{bmatrix} N_{dis}^{LAE} N_{dis}^{OUI} \end{bmatrix} \quad f_{OL}f_{DO} \begin{bmatrix} N_{dis}^{OUI} \left(N_{dis}^{OUI} - 1 \right) \right] \quad \left\{ f_{LL}f_{DO} + f_{DO}f_{LO} \right\} \begin{bmatrix} N_{dis}^{LAE} N_{dis}^{OUI} \end{bmatrix} \quad f_{OL}f_{DO$

where the second line follows from analogs of Equations $\overline{27,01}$ for the observed histograms. Similarly, we can take find the expectation value of (DR) histograms with Equation $\overline{13}$ as

$$\begin{split} \widehat{(DR)}_{A,gl}^{\text{obs}}(\boldsymbol{\delta}_{k}) &= \left(\sum_{1}^{N_{a}^{\text{obs}}} + \sum_{1}^{N_{a}^{\text{obs}}}\right) \sum_{j}^{N^{n}} \left[\langle u_{k}^{A} \Theta(\boldsymbol{\delta}_{lj} - \boldsymbol{\delta}_{\min,k}) | 1 - \Theta(\boldsymbol{\delta}_{lj} - \boldsymbol{\delta}_{\min,k}) | 1 \\ \Rightarrow \left\langle \widehat{(DR)}_{A,gl}^{\text{obs}}(\boldsymbol{\delta}_{k}) \right\rangle \approx \left(\sum_{1}^{N_{a}^{\text{obs}}} + \sum_{1}^{N_{a}^{\text{obs}}}\right) \sum_{j}^{N^{n}} \left[\langle u_{k}^{A} \rangle \Theta(\boldsymbol{\delta}_{lj} - \boldsymbol{\delta}_{\min,k}) | 1 - \Theta(\boldsymbol{\delta}_{lj} - \boldsymbol{\delta}_{\min,k}) | 1 \\ &= \left(\sum_{1}^{N_{a}^{\text{obs}}} \sum_{j}^{N^{n}} \langle (\boldsymbol{u}_{-A,l}) + \sum_{1}^{N_{a}^{\text{obs}}} \sum_{j}^{N^{n}} \langle (\boldsymbol{u}_{0,A,l}) \rangle \right) \Theta(\boldsymbol{\delta}_{lj} - \boldsymbol{\delta}_{\min,k}) | 1 - \Theta(\boldsymbol{\delta}_{lj} - \boldsymbol{\delta}_{\max,k}) | 1 \\ &= \sum_{1}^{N_{a}^{\text{obs}}} \sum_{j}^{N^{n}} \langle (\boldsymbol{u}_{-A,l}) + \sum_{1}^{N_{a}^{\text{obs}}} \sum_{j}^{N^{n}} \langle (\boldsymbol{u}_{0,A,l}) \rangle \right) \Theta(\boldsymbol{\delta}_{lj} - \boldsymbol{\delta}_{\min,k}) | 1 - \Theta(\boldsymbol{\delta}_{lj} - \boldsymbol{\delta}_{\max,k}) | 1 \\ &= \sum_{1}^{N_{a}^{\text{obs}}} \sum_{j}^{N^{n}} \langle (\boldsymbol{u}_{-A,l}) \otimes (\boldsymbol{\delta}_{lj} - \boldsymbol{\delta}_{\min,k}) | 1 - \Theta(\boldsymbol{\delta}_{lj} - \boldsymbol{\delta}_{\max,k}) | 1 \\ &= \sum_{1}^{N_{a}^{\text{obs}}} \sum_{j}^{N^{n}} \langle (\boldsymbol{u}_{-A,l}) \otimes (\boldsymbol{\delta}_{lj} - \boldsymbol{\delta}_{\min,k}) | 1 - \Theta(\boldsymbol{\delta}_{lj} - \boldsymbol{\delta}_{\max,k}) | 1 \\ &= \left(q_{l-A,l} \right) (DR)_{l}^{\text{obs}} N_{de}^{\text{obs}} N^{A} + \langle q_{D,A,l} \rangle (DR)_{l0}^{\text{obs}} N_{de}^{\text{obs}} N^{A} \\ &= \left[N_{a}^{\text{obs}} N^{A} \otimes (\boldsymbol{d}_{n-A,l}) N^{B} \langle (\boldsymbol{u}_{-A,l}) \right] \begin{bmatrix} (DR)_{l0}^{\text{obs}} N_{de}^{\text{obs}} N^{A} \\ (DR)_{l0}^{\text{obs}} \rangle \\ &= \left[N_{d}^{\text{obs}} N^{A} \otimes (\boldsymbol{d}_{n-A,l}) N^{B} \langle \boldsymbol{d}_{n-A,l} \rangle \right] \begin{bmatrix} (DR)_{l0}^{\text{obs}} N_{de}^{\text{obs}} N^{A} \\ (DR)_{l0}^{\text{obs}} \rangle \end{bmatrix} \\ &= \left[N_{d}^{2} N^{A} \otimes (\boldsymbol{d}_{n-A,l}) N^{B} \langle \boldsymbol{d}_{n-A,l} \rangle \right] \begin{bmatrix} (DR)_{l0}^{\text{obs}} N^{A} \otimes \boldsymbol{d}_{n-A,l} \rangle \\ (DR)_{l0}^{\text{obs}} \rangle \end{bmatrix} \\ &= \left[N_{d}^{2} N^{A} \otimes (\boldsymbol{d}_{n-A,l}) N^{B} \langle \boldsymbol{d}_{n-A,l} \rangle \right] \end{bmatrix}$$

$$= \left(\sum_{1}^{n}\sum_{j\neq n}^{n}\sum_{j\neq n}^{n}\sum_{j=1}^{n}\sum_{j=$$

Approximate the probabilities with their Expectation Values

In order to arrive at expressions that are comparable to Equations [27] Eq. i the weights with the respective averages, i.e., consider w_{A}^{*} as $\{w_{A}^{*}\}$. In this related $[q_{AB,A}] = f_{AB}$ (where i runs over M_{AB}^{*} objects) and $[g_{AB,A}] = f_{AB}$ (where i runs over M_{AB}^{*} objects) and $[g_{AB,A}] = f_{AB}$ + over M_{AB}^{*} objects). To first order, we ignore $\sigma_{BA,A}^{*}$ and $\sigma_{BA,A}^{*}$.

In order to write the expressions in terms of gs, we consider

$$\begin{split} \mathbf{g}_{LL} &= \left(p_{LL,s}\right) = \frac{N_{LL}^{L,dE}}{N_{LL}^{L,dF}} \left(\mathbf{g}_{LL,s}\right) = \frac{N_{LL}^{L,dE}}{N_{LL}^{L,dE}} f_{LL} \\ \mathbf{g}_{OO} &= \left(p_{OO(s)}\right) = \frac{N_{QL}^{QL'}}{N_{LL}^{QL'}} \left(\mathbf{g}_{OO(s)}\right) = \frac{N_{QL'}^{QL'}}{N_{LL}^{QL'}} f_{OO} \end{split}$$

$$\begin{split} & \left[\widehat{DD} \right]_{M,R,R}^{\infty}(\tilde{B}_{k}) = \left(\sum_{k=1}^{N} \sum_{j=1}^{N} \sum_{k=1}^{N} \sum_{j=1}^{N} \sum_{j=1}^{$$

Humna Awan

(DD)

(DD)

Probability-Weighted Estimator: De-Biasing

After all the algebra and some simplifications, we have



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

<u>Test</u>

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 <u>probabilistic</u> 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.

Humna Awan

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.
 - <u>Have the infrastructure to test different weights.</u>

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 RDI² 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 $w(\theta)$

Landy-Szalay estimator:

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

DD, DR, RR are histograms.

Explicitly, e.g.,

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

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

Galaxy Clustering: Traditional Estimator

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

• 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
⇒ 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.

Humna Awan