Controlling Systematics in Large-Scale Structure Surveys

Noah J. Weaverdyck
University of Michigan
nweaverd@umich.edu

Jan 26, 2021

Lawrence Berkeley National Laboratory

UNIVERSITY OF MICHIGAN

Outline

- Background
- Mitigation methods: insights from a common framework
- Simulated comparison
- Outlook

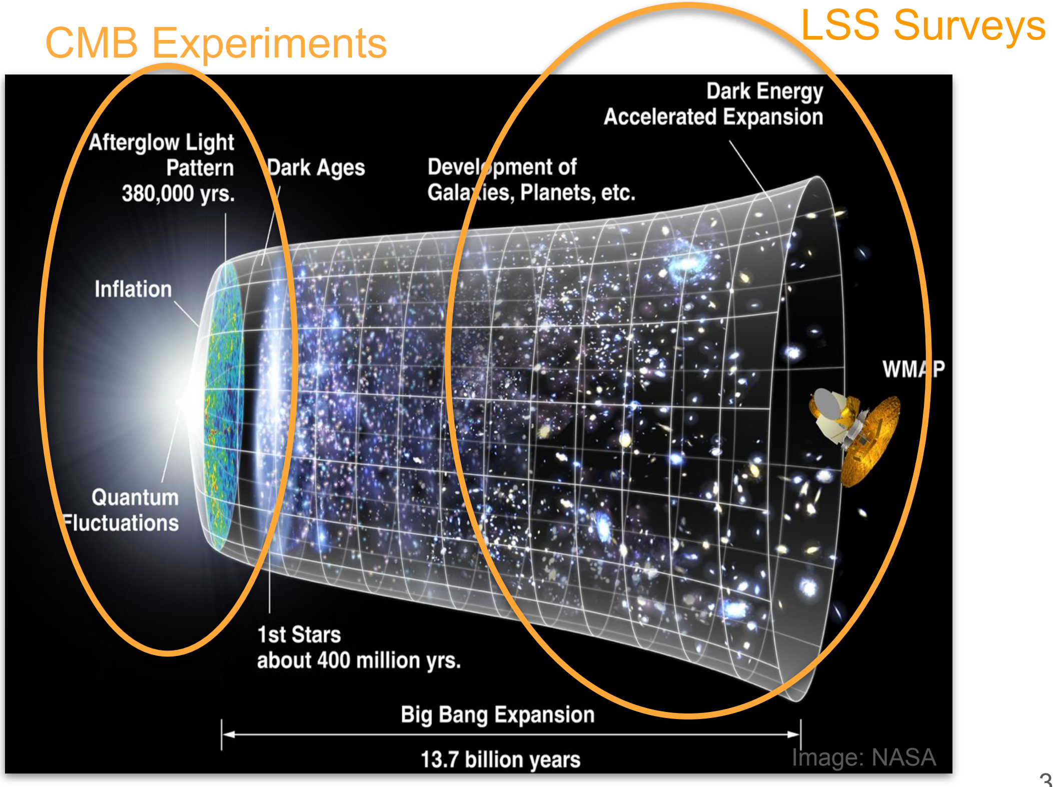
Largely based on
Weaverdyck & Huterer (2007.14499)

Some other work not covered in this talk:

- Rapid and generic systematics testing via importance sampling
- Small-scale modeling challenges for constraining inflation via the spectral runnings
- Vetting MCMC samplers for cosmological inference and model testing

Large-scale structure (LSS) surveys

- Map “late-time” density fluctuations
- Complement primordial fluctuations from CMB
- Probe *expansion history* and *growth of structure*;
  dark energy, neutrino mass, primordial non-Gaussianity

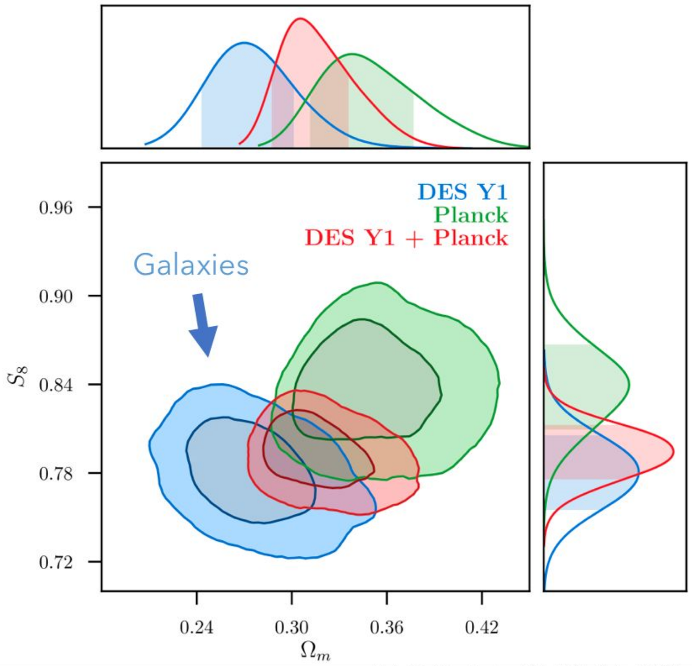


LSS surveys

- Primary observables:
  - Galaxy number density → galaxy clustering
  - Galaxy shapes → weak lensing

- Now competitive with CMB

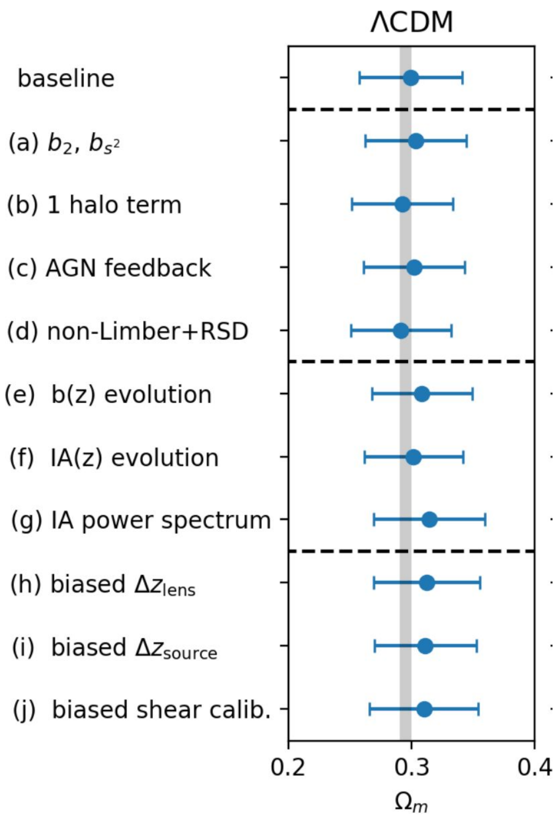
- LSST, DESI, Roman, SphereX...
  Large number densities → **small** statistical error
  - Control of systematics paramount to discover new physics



DES Collab (1708.01530)

LSS systematics

- Galaxy bias
- Small-scale modeling (non-linear Pk)
- Intrinsic alignments
- Photo-z errors
- ***Spatial systematics***
  - Modify selection function: map-level

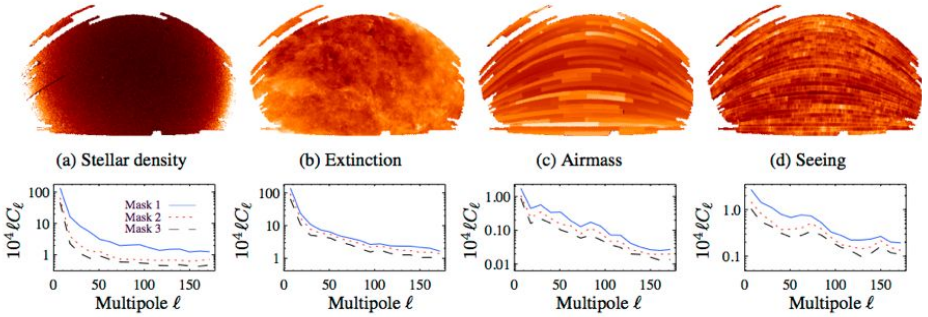


Krause et al. (DES) 1706.09359

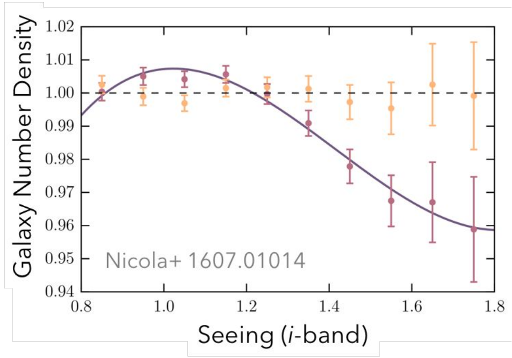
Spatial systematics

Observed galaxy field ≠ truth

- Astrophysical (stellar contamination, dust extinction, ...)
- Observing conditions (seeing, sky brightness, ...),
- Instrumental (flux calibration, source detection algorithms, ...)



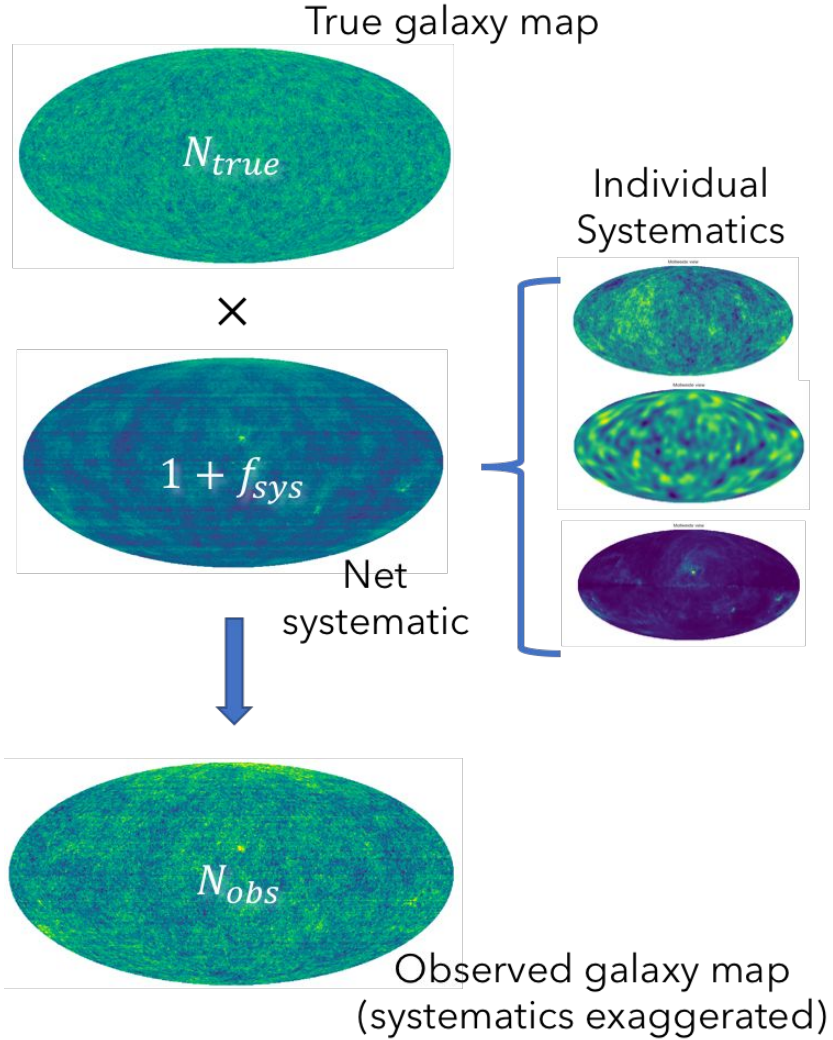
Leistedt et al. 1306.0005



Spatial systematics

- Spatially dependent screen (f_{sys}) modulates observed galaxy density

- Result: density maps biased!
(and 2-pt functions, 3-pt, …)



Spatial systematics: impact on ISW

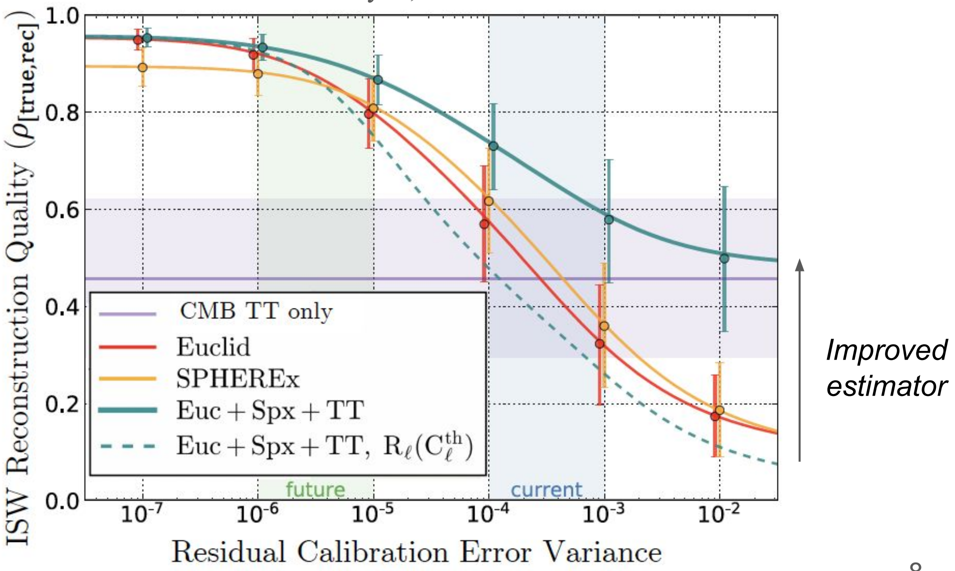
- First PhD project:
  assess control needed for accurate Integrated Sachs Wolfe effect (ISW)

- Leading contribution to CMB at large scales, important for DE/MG

- Infer from x-correlation with LSS

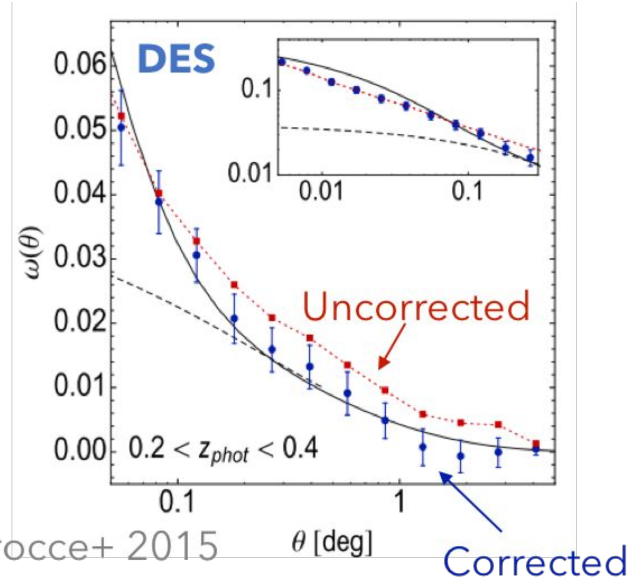
- Optimization, improved estimator for upcoming LSS surveys

Weaverdyck, Muir and Huterer 1709:08661



How to control systematics?

- Most common: use *systematic templates*, which trace potential contamination
  - Mask extreme regions
  - Estimate and correct for contamination (also: Balrog, Obiwan)
- Effects can be large
  - E.g. ELG and QSO densities in DESI imaging: ~10% variation after aggressive masking (Kitanidis et al. 1911.05714)
- Approaches varied, mostly ad hoc
  - Weaverdyck & Huterer (2007.14499): compare common methods, establish interpretive framework, improvements



Crocce+ 2015

How to control systematics?

Prominent methods investigated:

- Mode (De)Projection (e.g. HSC, SDSS QSOs)
- Multiple Linear Regression (e.g. KiDS LRGs, CFHTLenS)
- Template Subtraction (e.g. BOSS LRGs)
- DES-Y1 weighting (DES LRGs)
- "E.Net"
- "Forward Selection"

} New

All be reformulated as forms of **regression**

Mode (De)Projection

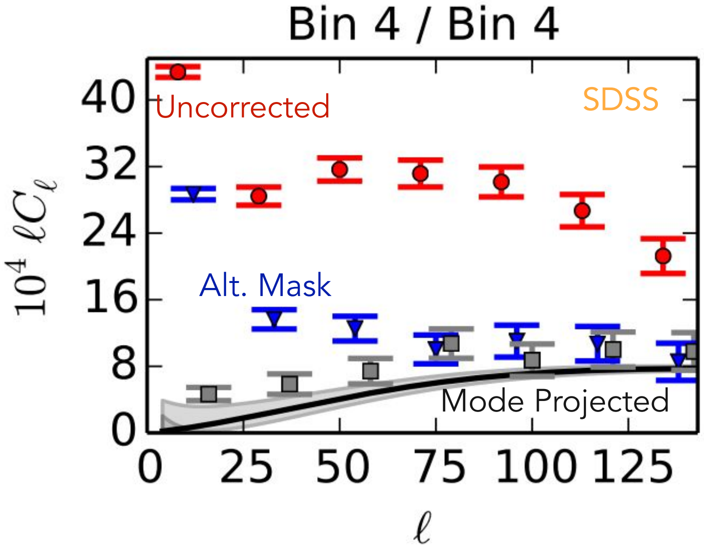
$$\delta_{\rm obs} \approx \delta_{\rm true} + \alpha t$$

Observed overdensity map

Template map

- Template map *t:*
  Marginalize over ***additive*** contaminant to overdensity
- *pseudo*-Cl version developed by Elsner+ 2016
  - Avoids N_{pix} x N_{pix} inversion
- Expanded to spin-2 fields
  - Public code *NaMaster* for LSST (Alonso+ 2018)
- Equivalent to OLS regression + step to debias *Cl*

Leistedt+ 2015



Mode (De)Projection

Template map

$$\delta_{\rm obs} \approx \delta_{\rm true} + \alpha t$$

Multiple systematic templates:

$t \to T$ $(\mathrm{N_{pix}} \times \mathrm{N_{tpl}})$

MP for Pseudo-Cl

$$\hat{\delta} = \boldsymbol{F}\delta_{\rm obs}$$

$$= \left[\lim_{\beta\to\infty} \left(I + \beta t t^{\dagger}\right)^{-1}\right]\delta_{\rm obs}$$

$$= \left[I - t(t^{\dagger}t)^{-1}t^{\dagger}\right]\delta_{\rm obs}$$

Map estimate

$$\hat{\delta} = \delta_{\rm obs} - t\hat{\alpha}$$

$$y = X\beta + \epsilon$$

$$\hat{\beta} = (X^{\dagger}X)^{-1}X^{\dagger}y$$

OLS to predict y from X

$$y = X(X^{\dagger}X)^{-1}X^{\dagger}y + \hat{\epsilon}$$

$$\delta_{\rm obs} = T[T^{\dagger}T]^{-1}T^{\dagger}\delta_{\rm obs} + \hat{\delta}$$

$\hat{\alpha}$ (for $[T^{\dagger}T]^{-1}T^{\dagger}\delta_{\rm obs}$)

Actually care about residuals and their clustering

MP estimate of contamination coefficient α
Is MLE, *assuming*:

$$\delta \sim \mathcal{N}(0, \sigma^2 I)$$

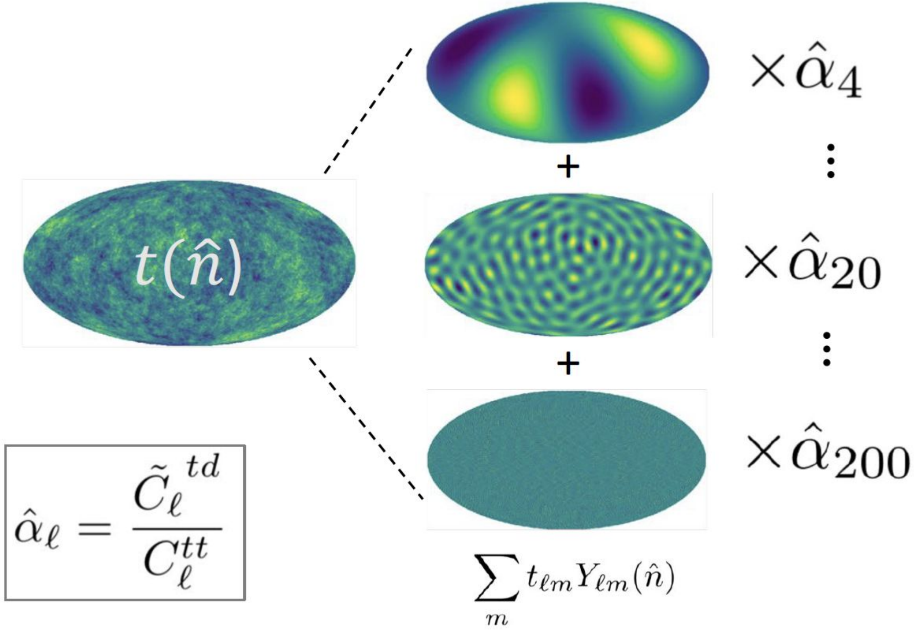
i.e. $$\hat{\alpha} = \mathrm{argmin}_{\alpha}||\delta_{\rm obs} - T\alpha||^2$$

(Ho+ 2011, with variants in Ross+ 2011, Crocce+ 2015, Kalus+ 2016)

$$\delta_{\mathrm{obs}} \approx \delta_{\mathrm{true}} + \alpha t$$

Template Subtraction

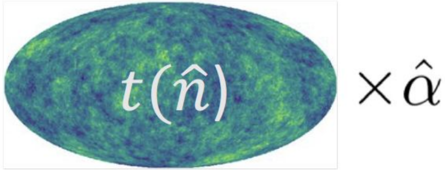
Decompose template, fit each harmonic



$$\hat{\alpha}_\ell = \frac{\tilde{C}_\ell^{td}}{C_\ell^{tt}}$$

(PCL) Mode Projection

Fit full template map



$$\hat{\alpha} = \frac{\sum_\ell (2\ell+1)\tilde{C}_\ell^{td}}{\sum_\ell (2\ell+1) C_\ell^{tt}}$$

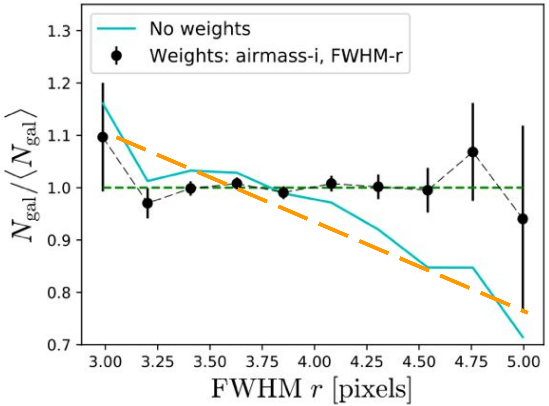
“Weights” method (DES-Y1)

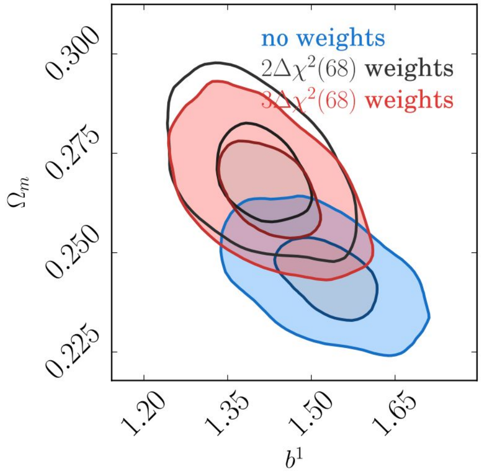
(Rodriquez-Monroy+ (in prep)
Elvin-Poole+ 2018, Ross+ 2011)

- Series of 1D, binned regressions on each template, iteratively reweight galaxies

- Pros vs OLS methods:
  - **Covariance** from mocks,
  - **Significance** threshold to control overfitting

- Cons vs OLS methods:
  - Only detect marginal relationships
  - Computation and time intensive (~1 day)





Elastic Net Weighting

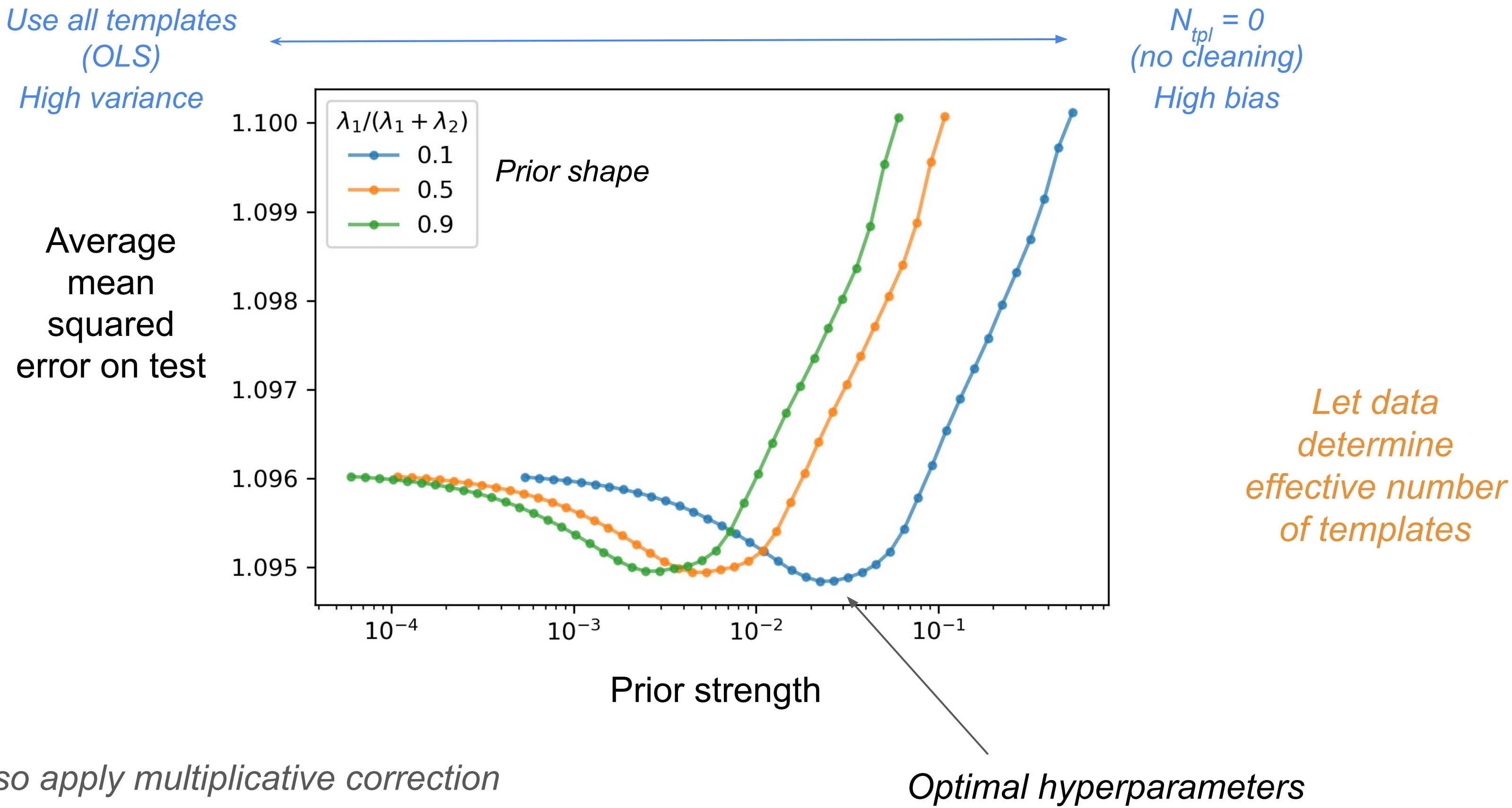
- Regression extension: form of regularization (Zou & Hastie 2005)
- Incorporate template selection, operate in full-D space

$$\hat{\alpha} = \text{argmin}_\alpha \left(||\delta_{\text{obs}} - T\alpha||^2 + \lambda_1 ||\alpha||_1 + \lambda_2 ||\alpha||_2^2 \right)$$

	$\lvert\lvert\delta_{\text{obs}} - T\alpha\rvert\rvert^2$	$\lambda_1 \lvert\lvert\alpha\rvert\rvert_1$	$\lambda_2 \lvert\lvert\alpha\rvert\rvert_2^2$
	OLS penalty	Sparsity prior (LASSO)	Regularization (Ridge)
In terms of Maximum Posterior Estimate, equivalent to:	*Gaussian Likelihood*	*Laplace prior on coefficients*	*Gaussian prior on coefficients*

In practice, select $\{\lambda_1,\lambda_2\}$ through cross-validation
(trained on subsets of the data)

Elastic Net Weighting



Also apply multiplicative correction

Multiplicative Correction

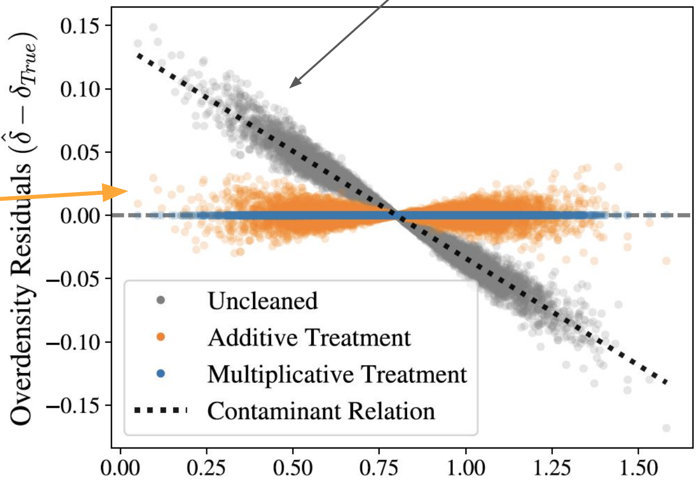
$$1 + \delta_{\rm obs} = (1 + \delta_{\rm true})(1 + f_{\rm sys})\gamma$$

$$\delta_{\rm obs} \approx \underline{\delta_{\rm true} + f_{\rm sys}} + \underline{\delta_{\rm true} f_{\rm sys}}$$

- Additive estimates (MP, EN, OLS...) leave residual scatter in map
  - Contaminant to small-scale power
- Remove with simple multiplicative correction

$$\hat{\delta} = \frac{\delta_{\rm obs} - \hat{f}_{\rm sys}}{1 + \hat{f}_{\rm sys}}$$

Each point = map pixel

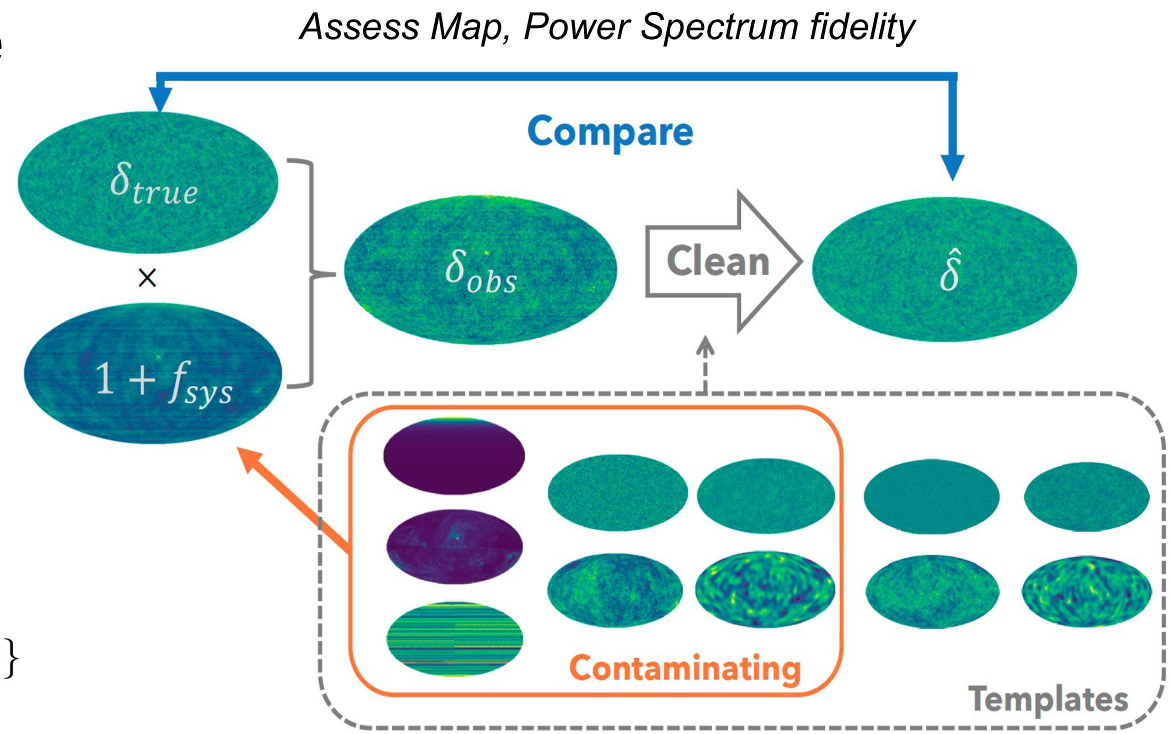


Seeing

Next → compare methods on simulation

Simulation Pipeline

- DES-Y5
- 5 z-bins
- Results not strongly sensitive to survey specs



Templates:

- Gaussian realizations

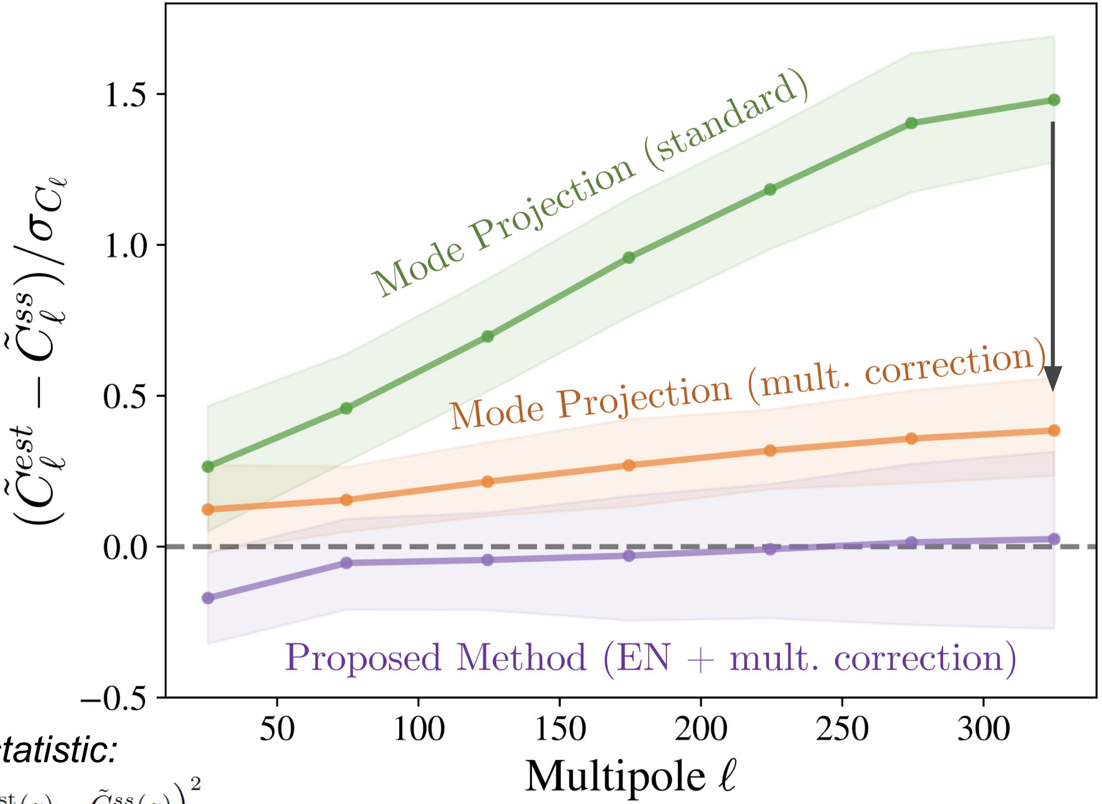
$$C_\ell \propto (\ell + 1)^{-p} \qquad p \in \{0, 1, 2\}$$

- Static (Dust, scanning strategy, etc)

Note: Methods applicable to any contaminated signal with templates. Here galaxy clustering, with signal = galaxy overdensity.
Generically: $\delta_{true} \rightarrow s$, $\delta_{obs} \rightarrow d_{obs}$

Importance of Multiplicative Correction

Power Spectrum Error



Significant improvement from multiplicative correction

68% spread across mocks

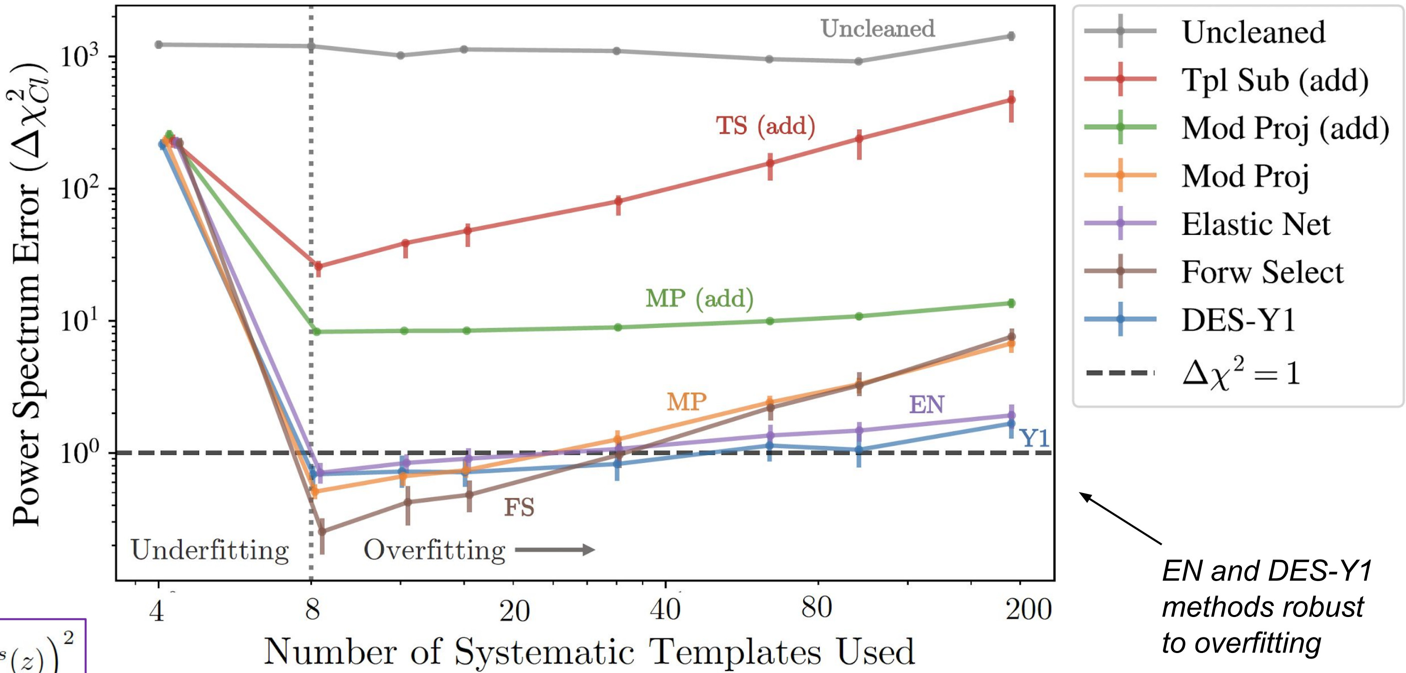
Quality summary statistic:

$$\Delta\chi^2_{C_\ell} = \sum_{z\text{bins}} \sum_{\ell=\ell_{\min}}^{350} \frac{\left(\tilde{C}_\ell^{\text{est}}(z) - \tilde{C}_\ell^{ss}(z)\right)^2}{\sigma^2_{C_\ell^{ss}(z)}},$$

Importance of Data-driven Template Selection

8 systematics

192 cleaning templates
= ~19 templates to 2nd order

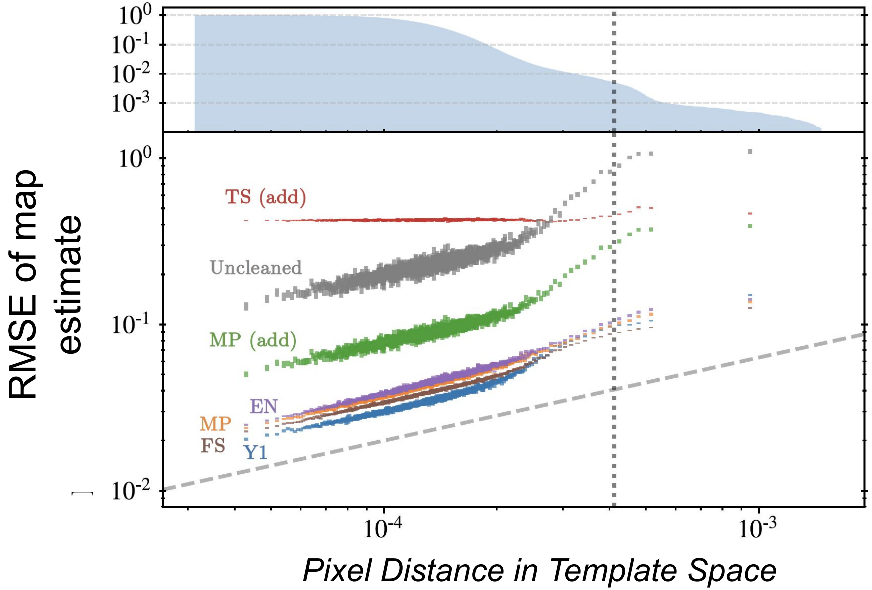


EN and DES-Y1 methods robust to overfitting

$$\Delta\chi^2_{C_\ell} = \sum_{z\text{bins}} \sum_{\ell=\ell_{\min}}^{350} \frac{\left(\tilde{C}_\ell^{\text{est}}(z) - \tilde{C}_\ell^{ss}(z)\right)^2}{\sigma^2_{C_\ell^{ss}(z)}},$$

Further development

- Mask optimization with template map-statistics
- Scale-optimized cleaning
  - Harmonic prewhitening
  - Maximize S/N for cosmology
- Systematics mitigation for primordial non-Gaussianity (f_{NL})
  - Key target of LSS
  - Cleaning large scales *crucial* (e.g. Castorina et al 2019)



Outlook

- Common framework unleashes new, powerful tools for systematics mitigation
  - Supervised learning/regression with *residuals* and clustering as signal of interest

- Corrections at *both* map and 2-pt function level

- Mask is important
  → rapid mitigation enables iteration

- Template selection should be *data-driven*
  - Self-calibrated sparsity + shrinkage priors work well!

Thank you!

R. Hahn

MP Assumptions on Noise

- True clustering signal = regression "noise"

  Only optimal if clustering signal

  1) Gaussian
  2) Diagonal
  3) Flat

  Can estimate α in pixel space or harmonic space $\hat{\alpha} = [T^\dagger T]^{-1} T^\dagger \delta_{\rm obs}$

	Pixel Space	**Harmonic Space**
Data	$\delta_{obs}(\hat{n}_i)$	$[\delta_{obs}]_{\ell m}$
Dims of T	$N_{pix} \times N_{tpl}$ (real)	$N_{\ell m} \times N_{tpl}$ (complex)
Regression Noise (additive)	$\delta(\hat{n}_i)$	$\delta_{\ell m}$
Gaussian	Approx. (~lognormal)	Yes
Diagonal	No	Yes
Flat	Yes	No

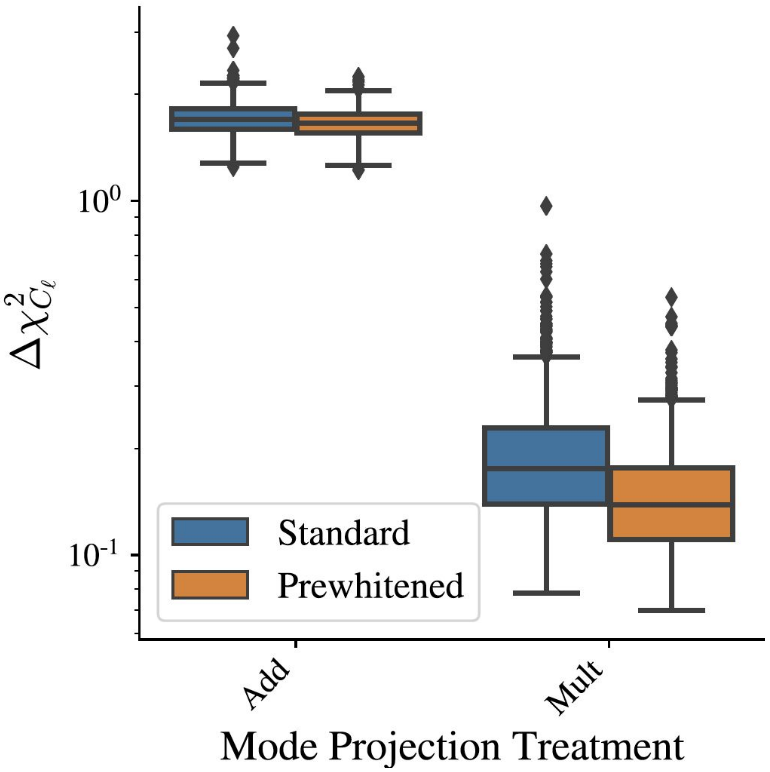
Diagonalize and optimally weight in harmonic space

$$\hat{\alpha} = \frac{\sum_{\ell=0}^{\ell_{\rm max}} (2\ell+1) \tilde{C}_\ell^{td} / C_\ell^{ss}}{\sum_{\ell=0}^{\ell_{\rm max}} (2\ell+1) \tilde{C}_\ell^{tt} / C_\ell^{ss}}.$$

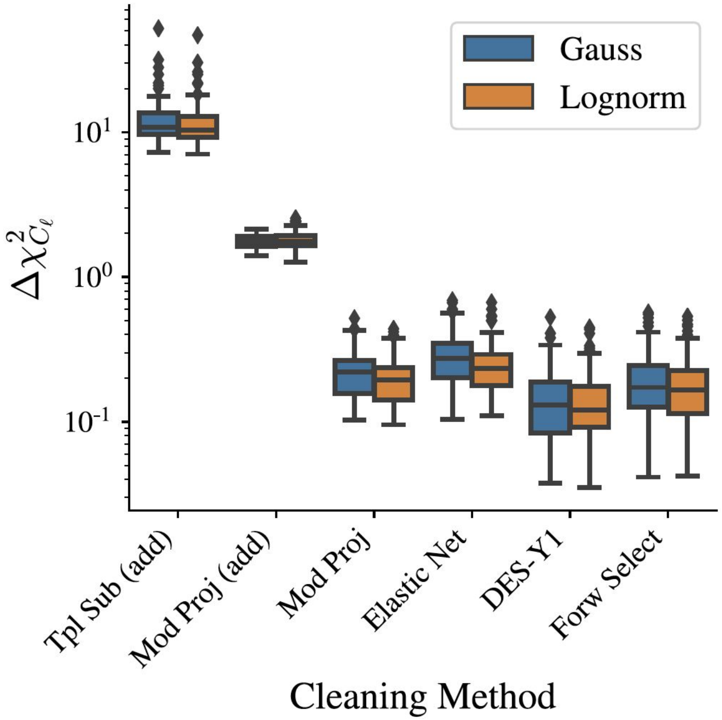
$$(d_{\rm obs})'_{\ell m} = (d_{\rm obs})_{\ell m} / \sqrt{C_\ell^{ss}}.$$

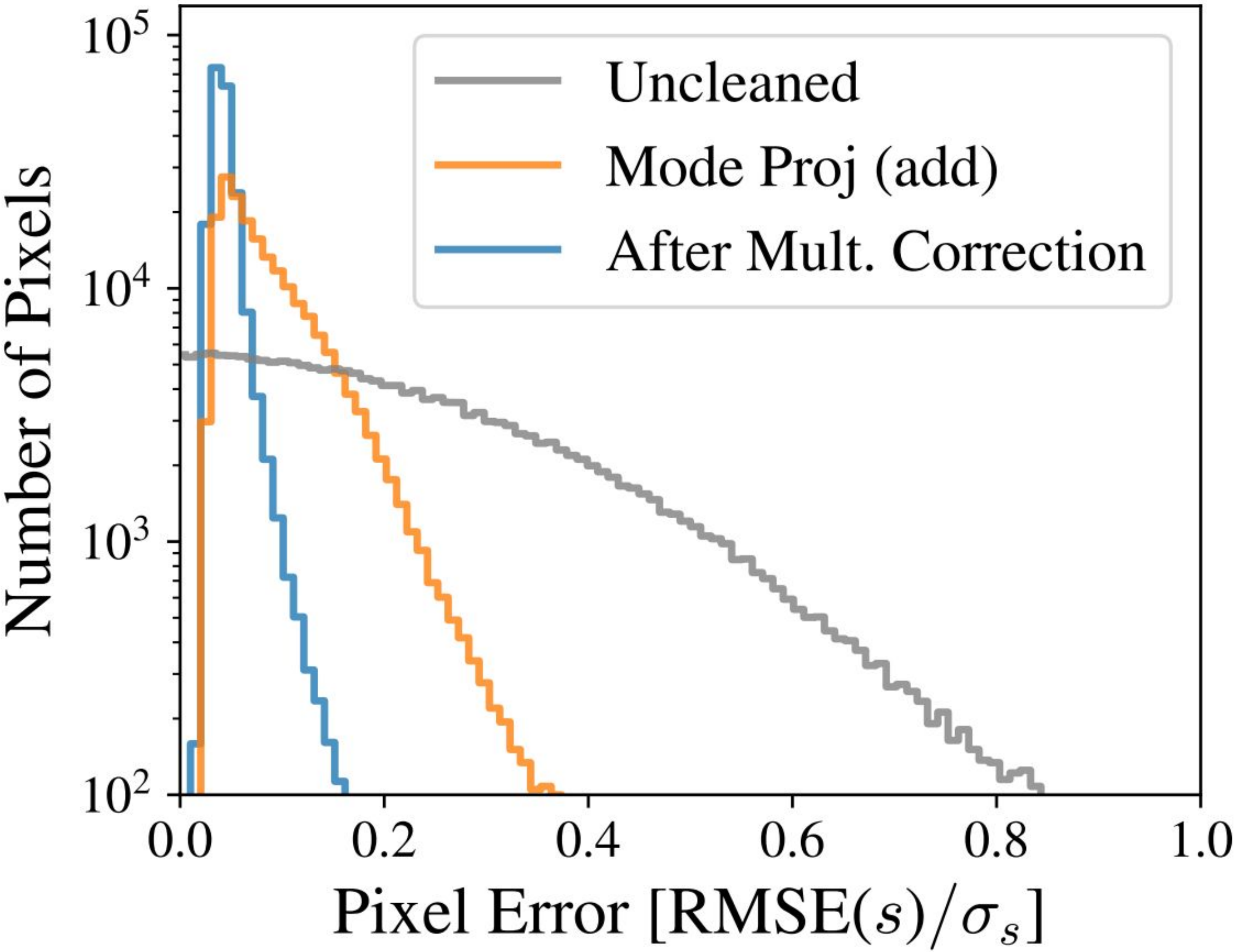
$$(t_i)'_{\ell m} = (t_i)_{\ell m} / \sqrt{C_\ell^{ss}},$$

Impact of pixel covariance
Minor compared to methodological differences.



No method particularly susceptible to Gaussian assumption





Uncleaned
Mode Proj (add)
After Mult. Correction
Number of Pixels
10^5
10^4
10^3
10^2
0.0
0.2
0.4
0.6
0.8
1.0
Pixel Error [RMSE(s)/σ_s]