Two insights from nonparametric statistics on cosmology research

Yen-Chi Chen

Department of Statistics
University of Washington

Nonparametric Statistics

- Nonparametric Statistics is a branch in statistics that attempts to make inference without using a parametric form of the underlying parameter of interest.
- Common topics: density estimation, regression, classification, clustering, ...

Nonparametric Statistics

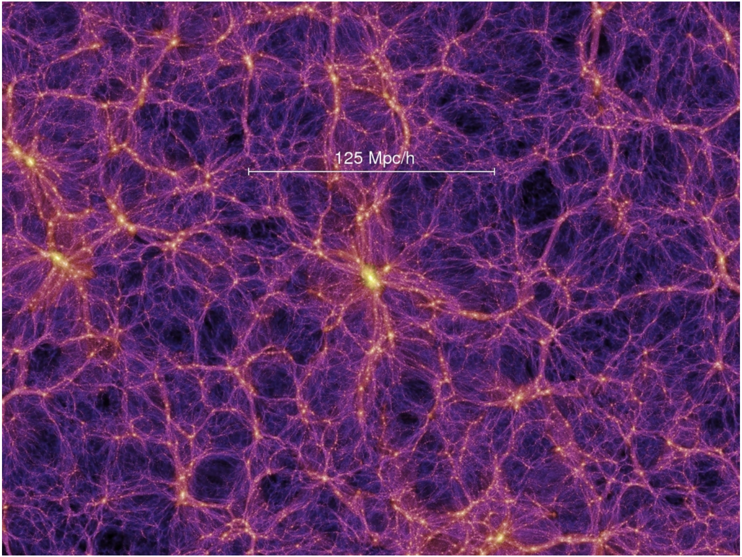
- Nonparametric Statistics is a branch in statistics that attempts to make inference without using a parametric form of the underlying parameter of interest.
- Common topics: density estimation, regression, classification, clustering, ...
- The spirit of nonparametrics also appears in other problem such as causal inference, graphical models, and the analysis of missing data (in particular, imputation).
- It offers a flexible way to investigate the underlying structures.

Nonparametric Statistics

- Nonparametric Statistics is a branch in statistics that attempts to make inference without using a parametric form of the underlying parameter of interest.
- Common topics: density estimation, regression, classification, clustering, ...
- The spirit of nonparametrics also appears in other problem such as causal inference, graphical models, and the analysis of missing data (in particular, imputation).
- It offers a flexible way to investigate the underlying structures.
- Examples in today's talk
  1. Density estimation and the discovery of large-scale structures.
  2. Analysis of bias from using the best fit (imputation).

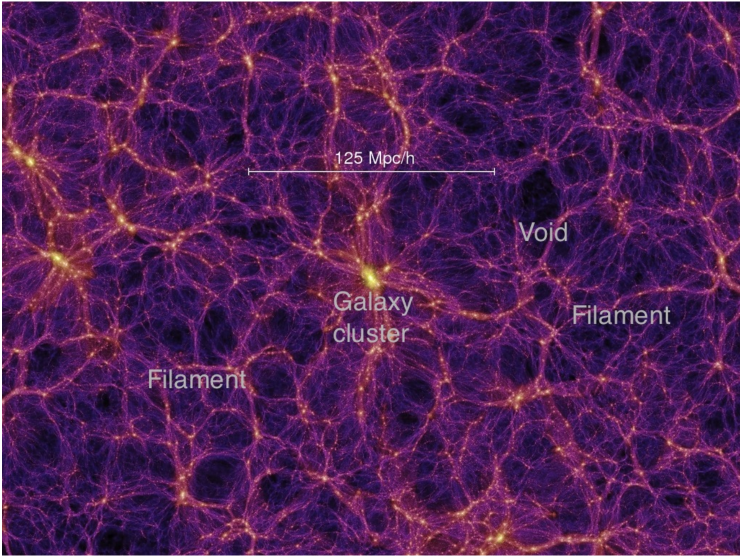
Part 1: Density estimation and detection of large-scale structures.

Cosmic Web: What Does Our Universe Look Like



Credit: Millennium Simulation

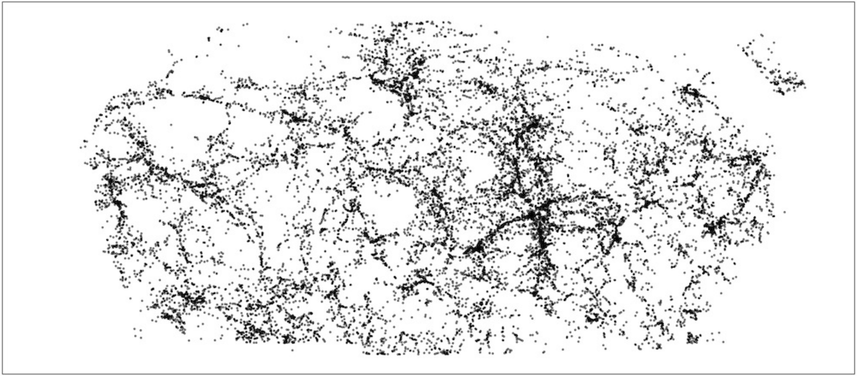
Cosmic Web: What Does Our Universe Look Like



Credit: Millennium Simulation

The Data

Here is what our data looks like:



Filament finding problem

- In simulations, we saw that there are clear filamentary structures.
- In the real data, we also saw some weakly filamentary forms in the distribution.
- How to recover filaments from the data is an open problem.

Consensus about Filaments

Different filament finders define filaments differently.
But there are some common properties that a filament should have (Bond et al. 1996):

Consensus about Filaments

Different filament finders define filaments differently.
But there are some common properties that a filament should have (Bond et al. 1996):

- It is a curve-like structure.

Consensus about Filaments

Different filament finders define filaments differently.
But there are some common properties that a filament should have (Bond et al. 1996):

- It is a curve-like structure.
- It characterizes high (matter) density area.

Consensus about Filaments

Different filament finders define filaments differently.
But there are some common properties that a filament should have (Bond et al. 1996):

- It is a curve-like structure.
- It characterizes high (matter) density area.
- It shows connectivity of the matter distribution.

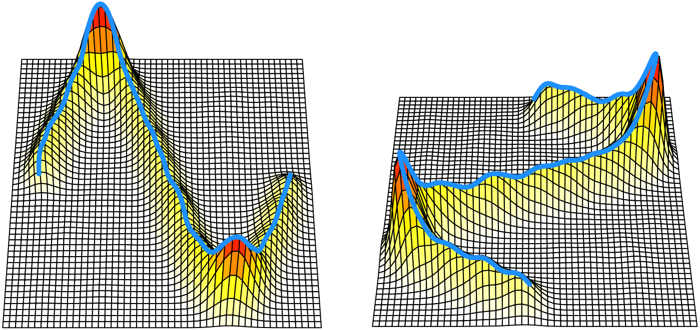
Density Ridges

We formalize the notion of filaments as *density ridges*.

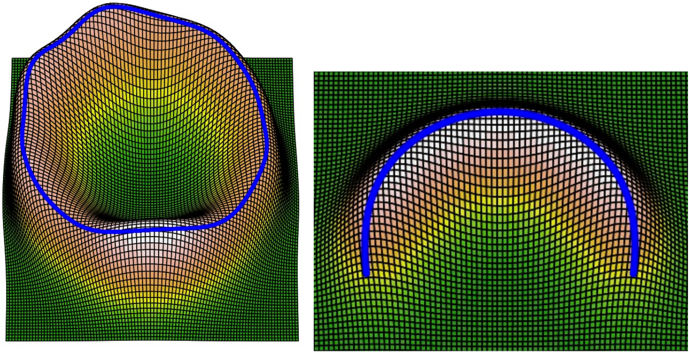
Example: Ridges in Mountains

Credit: Google

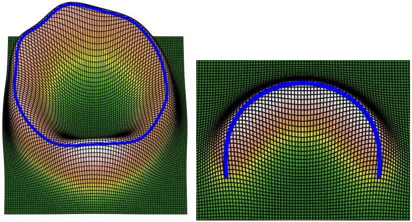
Example: Ridges in Smooth Functions



Example: Ridges in Smooth Functions

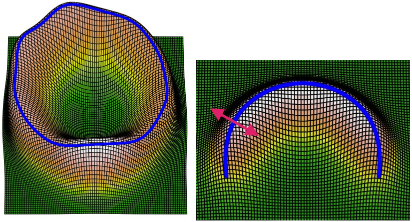


Ridges: Local Modes in Subspace



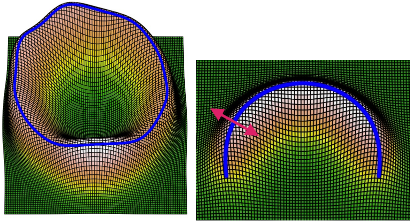
- A generalized local mode in a specific ‘subspace’.

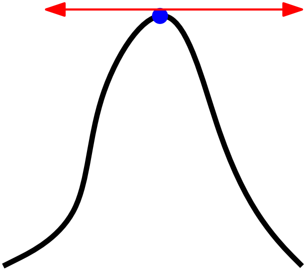
Ridges: Local Modes in Subspace



- A generalized local mode in a specific 'subspace'.

Ridges: Local Modes in Subspace

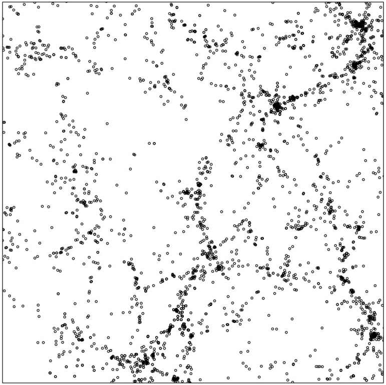




- A generalized local mode in a specific ‘subspace’.

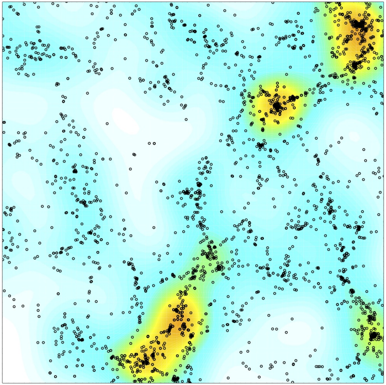
Finding ridges

- Original data.



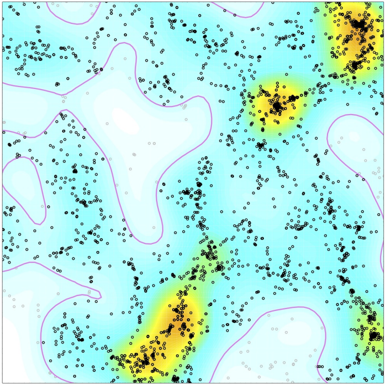
Finding ridges

- Original data.
- Density estimation.



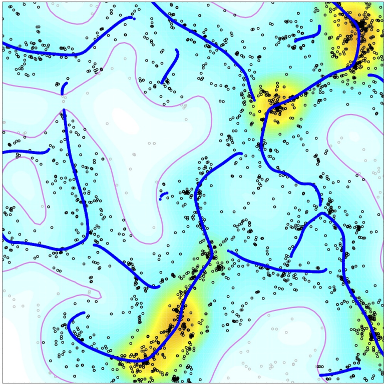
Finding ridges

- Original data.
- Density estimation.
- Thresholding (denoising).

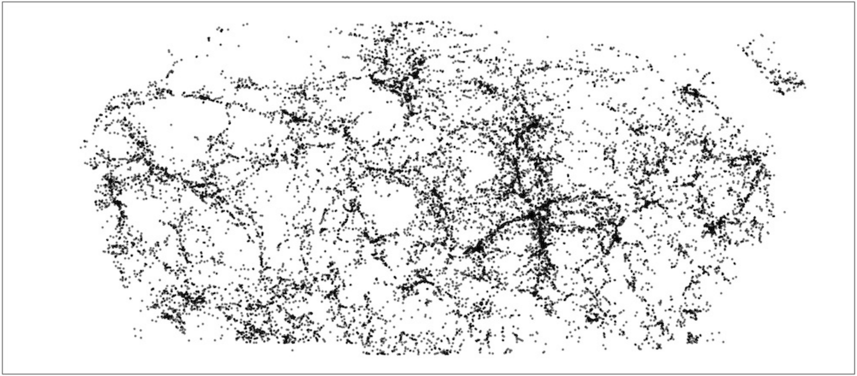


Finding ridges

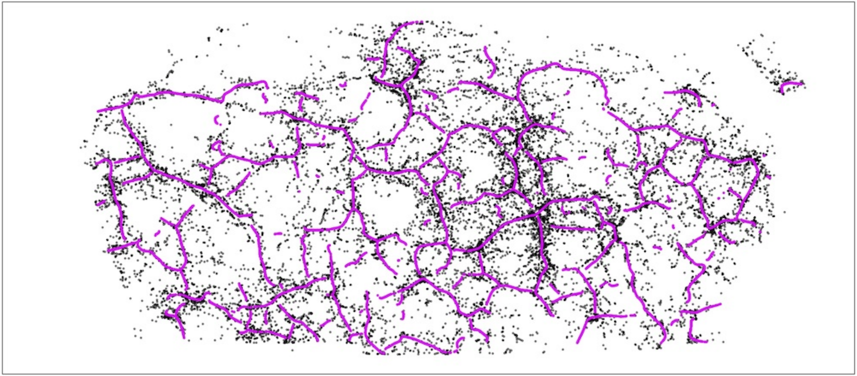
- Original data.
- Density estimation.
- Thresholding (denoising).
- Ridge finding.



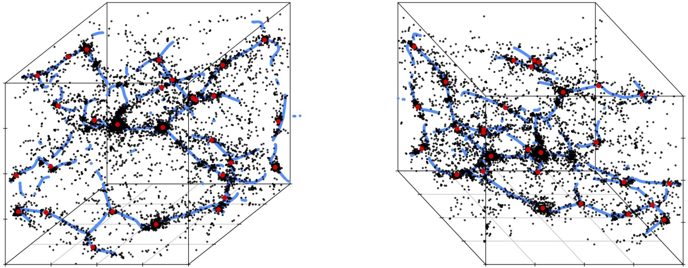
Example for Estimated Density Ridges



Example for Estimated Density Ridges



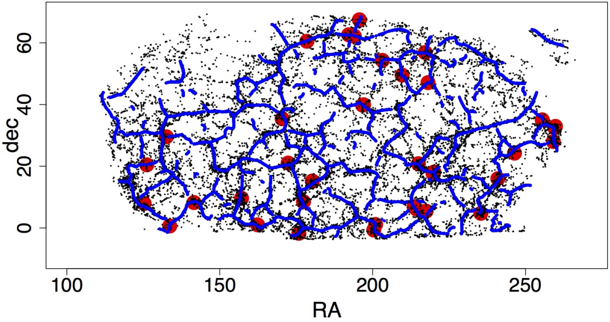
3D Example for Estimated Ridges



Blue curves: density ridges.
Red points: density local modes.

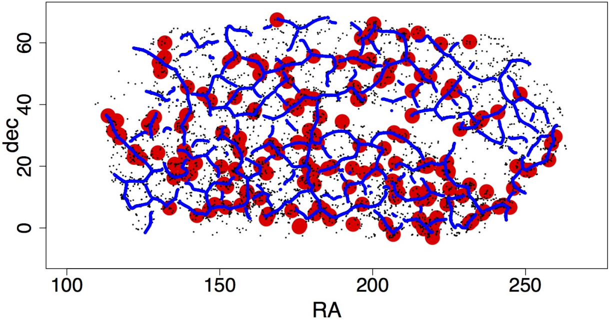
SDSS: Comparing to Clusters

- Blue: filaments. Red: galaxy clusters (redMaPPer).



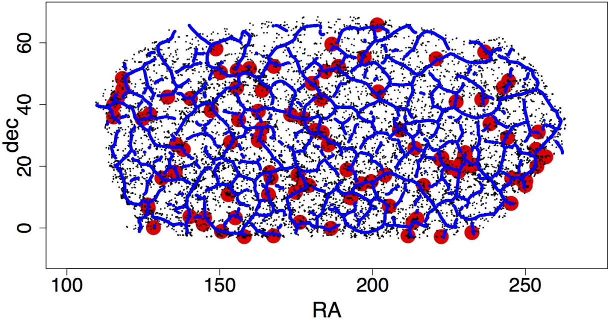
SDSS: Comparing to Clusters

- Blue: filaments. Red: galaxy clusters (redMaPPer).



SDSS: Comparing to Clusters

- Blue: filaments. Red: galaxy clusters (redMaPPer).

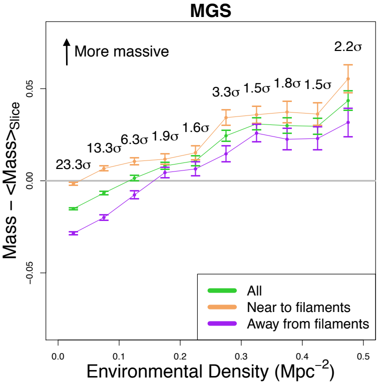


SDSS: Filament Effects VS Environments

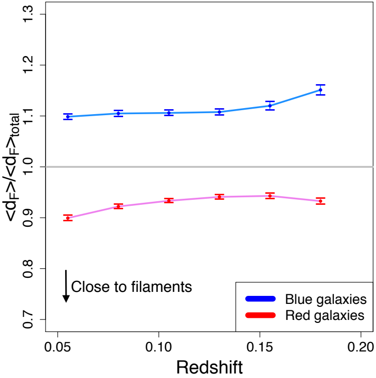
Do filaments have an extra effect other than environments?

SDSS: Filament Effects VS Environments

Do filaments have an extra effect other than environments?
$\longrightarrow$Yes!



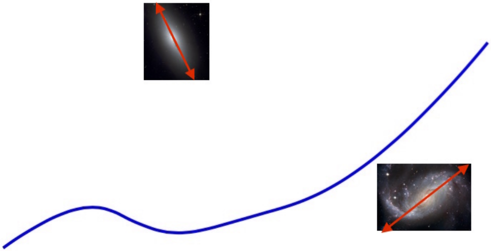
SDSS: Color



Similar pattern also appears for other galaxy properties such as brightness, size, and age.

The Alignment of a Galaxy along a Filament - 1

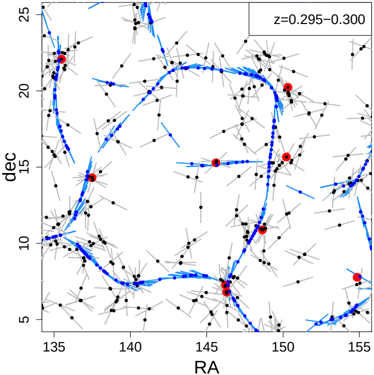
Theorists have conjectured about the alignment of galaxy along nearby filaments.



We now try to test such a conjecture.

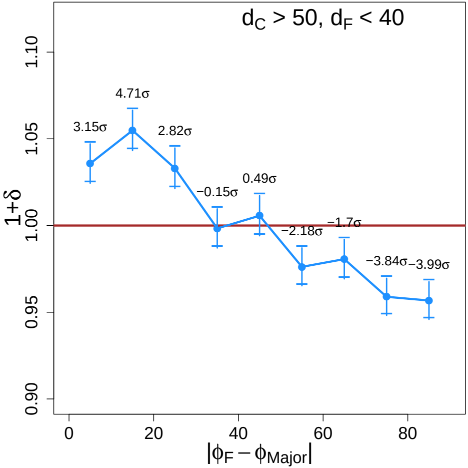
The Alignment of a Galaxy along a Filament - 2

We can easily define the orientation of filament because it is a curve.
For each galaxy, we measure its orientation by fitting an ellipse.



We are interested in the inner product between the major axis of a galaxy and the orientation of the nearest filament.

Excess Probability Density



Y-axis: the ratio of observed angular distribution versus a uniform distribution over $[0, 90]$ deg.
If no alignment, the ratio should be 1.

Part 2: The danger of using the best fit

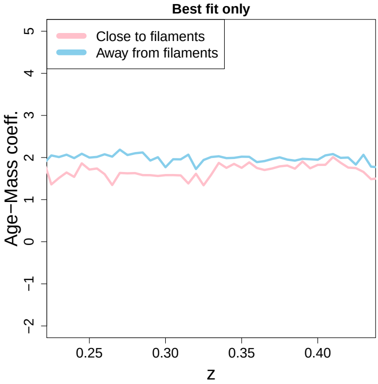
A common value-added data

	Mass	M. err	Age	A. err	RA	dec	redshift	others
Galaxy 1	M_1	$E_{M,1}$	A_1	$E_{A,1}$	RA_1	dec_1	Z_1	O_1
Galaxy 2	M_2	$E_{M,2}$	A_2	$E_{A,2}$	RA_2	dec_2	Z_2	O_2
Galaxy 3	M_3	$E_{M,3}$	A_3	$E_{A,3}$	RA_3	dec_3	Z_3	O_3
Galaxy 4	M_4	$E_{M,4}$	A_4	$E_{A,4}$	RA_4	dec_4	Z_4	O_4
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

- Blue variables: directly observed using the telescope.
- Red variables: unobserved, inferred from the observables.
- Q: is it reliable to use the inferred variables (often best fitted) to make scientific conclusion?

A motivating example: detecting filament effects

Here we attempted to analyze the effect from filaments on the age-mass relation (regression coefficient).
We use the best fitted mass and age from the data.



Model prediction

- For a galaxy, let O denotes all its observed profiles.
- A model that associates O with age (A) and mass (M) can be viewed as a distribution/likelihood of A, M given O :

$$p(A = a, M = m | O)$$

Model prediction

- For a galaxy, let O denotes all its observed profiles.
- A model that associates O with age (A) and mass (M) can be viewed as a distribution/likelihood of A, M given O :

$$p(A = a, M = m|O)$$

- In many data products, we have predicted mass and age of each galaxy. Generally, the predicted mass and age are

$$(\widehat{A}, \widehat{M}) = \text{argmax}_{A,M} p(A, M|O).$$

Namely, they are the best fitted values in the model.

Model prediction

- For a galaxy, let O denotes all its observed profiles.
- A model that associates O with age (A) and mass (M) can be viewed as a distribution/likelihood of A, M given O :

$$p(A = a, M = m|O)$$

- In many data products, we have predicted mass and age of each galaxy. Generally, the predicted mass and age are

$$(\widehat{A}, \widehat{M}) = \text{argmax}_{A,M} p(A, M|O).$$

  Namely, they are the best fitted values in the model.
- In our previous analysis, we were computing the association between age and mass via the best fitted value $\widehat{A}, \widehat{M}$.
- Namely, we are ignoring the uncertainty of A, M in our analysis. Will this be okay?

Examining simulations

- To investigate the effect of ignoring the uncertainty, we use simulation data.
- Here we use the MassiveBlack-II simulation. We know the true mass and age of each galaxy.

Examining simulations

- To investigate the effect of ignoring the uncertainty, we use simulation data.
- Here we use the MassiveBlack-II simulation. We know the true mass and age of each galaxy.
- To build a prediction model, we use the number density (here we use the distance to the 50th nearest neighbor) as the predictor.

Examining simulations

- To investigate the effect of ignoring the uncertainty, we use simulation data.
- Here we use the MassiveBlack-II simulation. We know the true mass and age of each galaxy.
- To build a prediction model, we use the number density (here we use the distance to the 50th nearest neighbor) as the predictor.
- So our simulation data can be summarized as

$$(M_1, A_1, O_1), \cdots, (M_n, A_n, O_n),$$

where M_i is the true mass A_i is the true age and O_i is the number density.

Examining simulations

- To investigate the effect of ignoring the uncertainty, we use simulation data.
- Here we use the MassiveBlack-II simulation. We know the true mass and age of each galaxy.
- To build a prediction model, we use the number density (here we use the distance to the 50th nearest neighbor) as the predictor.
- So our simulation data can be summarized as

$$(M_1, A_1, O_1), \cdots, (M_n, A_n, O_n),$$

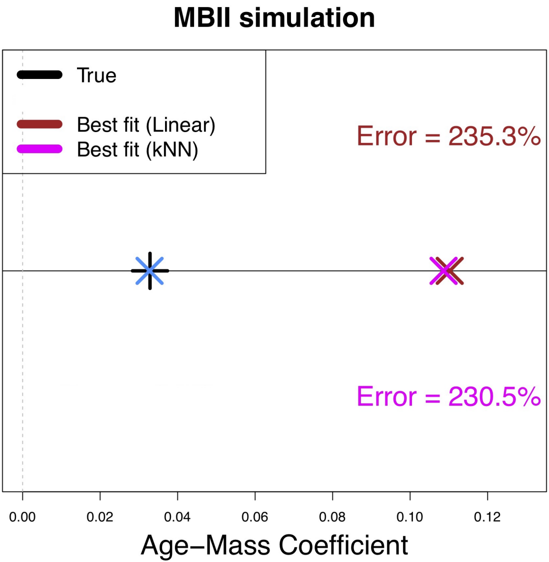
  where M_i is the true mass A_i is the true age and O_i is the number density.
- To predict M, A from O, we consider a simple linear regression and a 50-nearest neighbor (NN) regression.

Simulations: updated data

	Original		Linear model		50-NN	
Galaxy 1	M_1	A_1	$\widehat{M}_{LM,1}$	$\widehat{A}_{LM,1}$	$\widehat{M}_{kNN,1}$	$\widehat{A}_{kNN,1}$
Galaxy 2	M_2	A_2	$\widehat{M}_{LM,2}$	$\widehat{A}_{LM,2}$	$\widehat{M}_{kNN,2}$	$\widehat{A}_{kNN,2}$
Galaxy 3	M_3	A_3	$\widehat{M}_{LM,3}$	$\widehat{A}_{LM,3}$	$\widehat{M}_{kNN,3}$	$\widehat{A}_{kNN,3}$
Galaxy 4	M_4	A_4	$\widehat{M}_{LM,4}$	$\widehat{A}_{LM,4}$	$\widehat{M}_{kNN,4}$	$\widehat{A}_{kNN,4}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

- The true regression coefficient is obtained by regressing $Y = A_i$ with $X = M_i$.
- Question: if we use the best fitted/predicted values from the linear model or nonparametric model, will we obtain a similar regression coefficients?

Failure of best fitted method



No! It fails miserably!

Why the best fitted values fail?

- Often the best predicted value of A, M from O is the conditional mean $\mathbb{E}(A|O)$ and $\mathbb{E}(M|O)$.
- Regressing A with M is different from regressing $\mathbb{E}(A|O)$ with $\mathbb{E}(M|O)$!

Why the best fitted values fail?

- Often the best predicted value of A, M from O is the conditional mean $\mathbb{E}(A|O)$ and $\mathbb{E}(M|O)$.
- Regressing A with M is different from regressing $\mathbb{E}(A|O)$ with $\mathbb{E}(M|O)$!
- Take the covariance as an example, by law of total covariance,

$$\text{Cov}(A, M) = \text{Cov}(\mathbb{E}(A|O), \mathbb{E}(M|O)) + \mathbb{E}(\text{Cov}(\text{A}, \text{M}|\text{O})).$$

- The first term is what we compute when using the best fitted value.

Why the best fitted values fail?

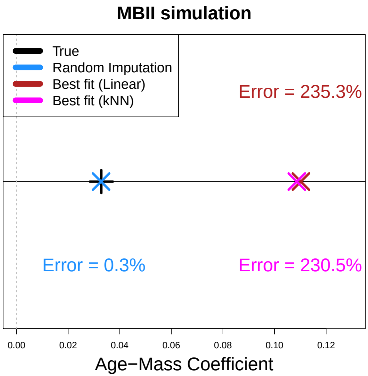
- Often the best predicted value of A, M from O is the conditional mean $\mathbb{E}(A|O)$ and $\mathbb{E}(M|O)$.
- Regressing A with M is different from regressing $\mathbb{E}(A|O)$ with $\mathbb{E}(M|O)$!
- Take the covariance as an example, by law of total covariance,

$$\mathrm{Cov}(A, M) = \mathrm{Cov}(\mathbb{E}(A|O), \mathbb{E}(M|O)) + \mathbb{E}(\mathrm{Cov}(\mathrm{A}, \mathrm{M}|\mathrm{O})).$$

- The first term is what we compute when using the best fitted value.
- But it ignores the second term!

A simple remedy: random imputation

- Here is a simple remedy: instead of using the best fitted value, we take a *random draw* from the conditional density $p(A, M|O)$ (known as random imputation)[1].



[1] Estimated by the 50-NN in this case.

Why random imputation works? - 1

- In the ideal case where we get to observe the age and mass directly, our data can be summarized as IID random vectors

$$(M_1, A_1, O_1), \cdots, (M_n, A_n, O_n) \sim p(m, a, o),$$

where $p(m, a, o)$ is the joint density of M, A, O.

Why random imputation works? - 1

- In the ideal case where we get to observe the age and mass directly, our data can be summarized as IID random vectors

$$(M_1, A_1, O_1), \cdots, (M_n, A_n, O_n) \sim p(m, a, o),$$

where $p(m, a, o)$ is the joint density of M, A, O.

- A measure of association between age and mass can often be written as $\theta(M, A)$ and we are interested in the population average

$$\theta = \mathbb{E}(\theta(M, A)) = \int \theta(m, a)p(m, a)dmda,$$

where $p(m, a)$ is the joint density of M, A.

Why random imputation works? - 1

- In the ideal case where we get to observe the age and mass directly, our data can be summarized as IID random vectors

$$(M_1, A_1, O_1), \cdots, (M_n, A_n, O_n) \sim p(m, a, o),$$

where $p(m, a, o)$ is the joint density of M, A, O.

- A measure of association between age and mass can often be written as $\theta(M, A)$ and we are interested in the population average

$$\theta = \mathbb{E}(\theta(M, A)) = \int \theta(m, a) p(m, a) dmda,$$

where $p(m, a)$ is the joint density of M, A.

- In practice, mass and age are missing, what we observe are IID random elements

$$O_1, \cdots, O_n \sim p(o).$$

Why random imputation works? - 2

- The decomposition

$$p(m, a, o) = p(m, a|o)p(o)$$

implies that if we augment the i-th galaxy with random numbers (M_i^*, A_i^*) from $p(m, a|O_i)$, the triplet can be viewed from

$$(M_i^*, A_i^*, O_i) \sim p(m, a|o)p(o) = p(m, a, o).$$

Why random imputation works? - 2

- The decomposition

$$p(m, a, o) = p(m, a|o)p(o)$$

implies that if we augment the i-th galaxy with random numbers (M_i^*, A_i^*) from $p(m, a|O_i)$, the triplet can be viewed from

$$(M_i^*, A_i^*, O_i) \sim p(m, a|o)p(o) = p(m, a, o).$$

- Thus, as long as we independently draw mass and age from the conditional density, we obtain a dataset that behaves like a fully observed data.

Why random imputation works? - 2

- The decomposition

$$p(m, a, o) = p(m, a|o)p(o)$$

implies that if we augment the i-th galaxy with random numbers (M_i^*, A_i^*) from $p(m, a|O_i)$, the triplet can be viewed from

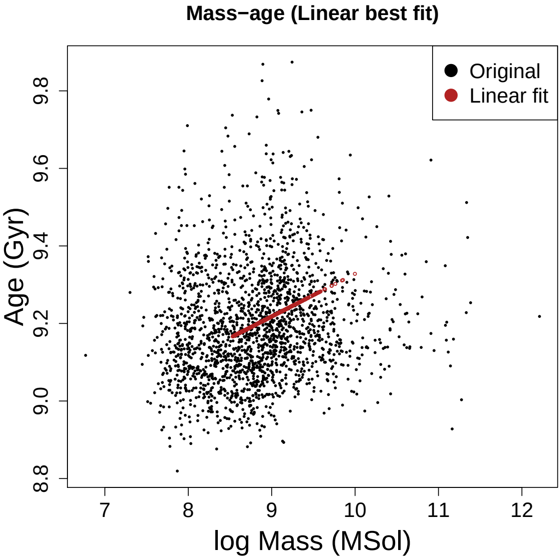
$$(M_i^*, A_i^*, O_i) \sim p(m, a|o)p(o) = p(m, a, o).$$

- Thus, as long as we independently draw mass and age from the conditional density, we obtain a dataset that behaves like a fully observed data.
- Then we can use

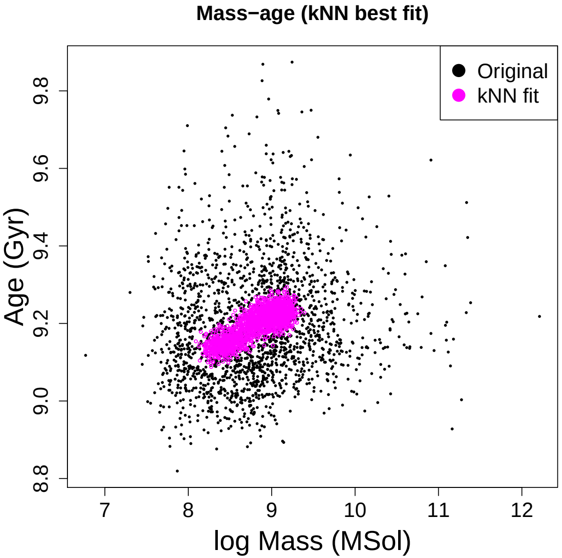
$$(M_1^*, A_1^*), \cdots, (M_n^*, A_n^*)$$

to accurately estimate $\theta = \mathbb{E}(\theta(M, A))$.

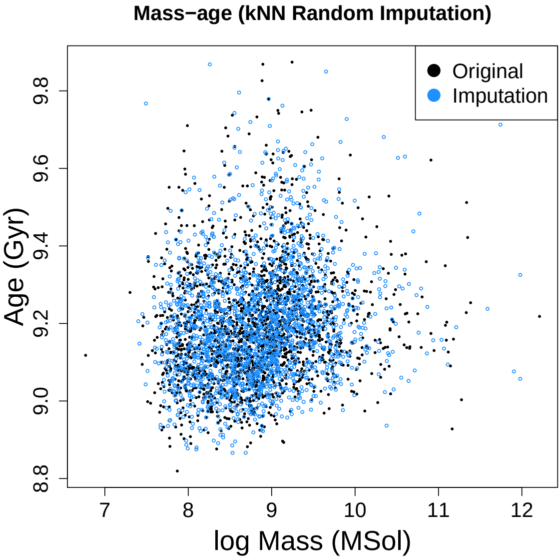
Why random imputation works? (Visually)



Why random imputation works? (Visually)



Why random imputation works? (Visually)



Multiple imputation and Monte Carlo errors

- Although the above procedure gives us an unbiased estimator, it may suffer from the Monte Carlo errors if we only impute the unobserved entries once.
- In general, we should repeat this imputation multiple times, creating multiple imputed data, and compute the final estimates.[2]

[2] known as the multiple imputation.

Multiple imputation and Monte Carlo errors

- Although the above procedure gives us an unbiased estimator, it may suffer from the Monte Carlo errors if we only impute the unobserved entries once.
- In general, we should repeat this imputation multiple times, creating multiple imputed data, and compute the final estimates.[2]
- Luckily, in most Astronomy survey, the sample size is large so the Monte Carlo errors are small.

[2] known as the multiple imputation.

What if we only know the marginal error? - 1

Recall the original dataset:

	Mass	M. err	Age	A. err	RA	dec	redshift	others
Galaxy 1	M_1	$E_{M,1}$	A_1	$E_{A,1}$	RA_1	dec_1	Z_1	O_1
Galaxy 2	M_2	$E_{M,2}$	A_2	$E_{A,2}$	RA_2	dec_2	Z_2	O_2
Galaxy 3	M_3	$E_{M,3}$	A_3	$E_{A,3}$	RA_3	dec_3	Z_3	O_3
Galaxy 4	M_4	$E_{M,4}$	A_4	$E_{A,4}$	RA_4	dec_4	Z_4	O_4
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

- We do have the errors that represents the marginal distribution of $p(M|O)$ and $p(A|O)$.
- If this is all we have, can we make a better inference?
- Note: $E_{M,1}$ can be viewed as the SD of $p(M|O)$.

What if we only know the marginal error? - 2

- If we assume that $M|O$ and $A|O$ follow a normal distribution, then the above table gives us information about $M_i|O_i$ and $A_i|O_i$:

$$M_i|O_i \sim N(M_i, E_{M,i}), \quad A_i|O_i \sim N(A_i, E_{A,i}).$$

- It seems that we can generate from the distribution $p(m, a|o)$ using this information.

What if we only know the marginal error? - 2

- If we assume that $M|O$ and $A|O$ follow a normal distribution, then the above table gives us information about $M_i|O_i$ and $A_i|O_i$:

$$M_i|O_i \sim N(M_i, E_{M,i}), \quad A_i|O_i \sim N(A_i, E_{A,i}).$$

- It seems that we can generate from the distribution $p(m, a|o)$ using this information.
- Actually, we CANNOT–we still need to know the (conditional) correlation between the two random variables.

What if we only know the marginal error? - 2

- If we assume that $M|O$ and $A|O$ follow a normal distribution, then the above table gives us information about $M_i|O_i$ and $A_i|O_i$:

$$M_i|O_i \sim N(M_i, E_{M,i}), \quad A_i|O_i \sim N(A_i, E_{A,i}).$$

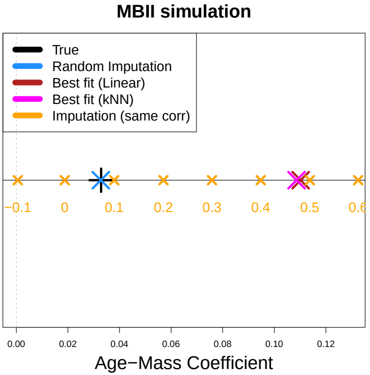
- It seems that we can generate from the distribution $p(m, a|o)$ using this information.
- Actually, we CANNOT–we still need to know the (conditional) correlation between the two random variables.
- Namely, we need $\text{Cor}(A_i, M_i|O_i)$ to reconstruct $p(a, m|o)$.

Sensitivity analysis: a partial solution

- Here is a simple method to roughly investigate the effect–we assume a single number for all correlations and evaluate how it influences our final result.

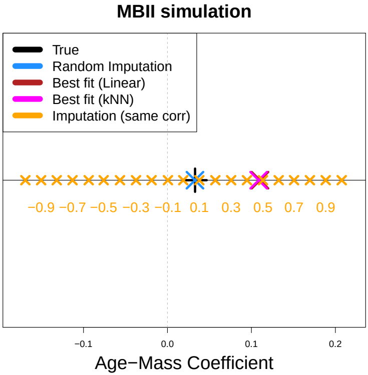
Sensitivity analysis: a partial solution

- Here is a simple method to roughly investigate the effect–we assume a single number for all correlations and evaluate how it influences our final result.



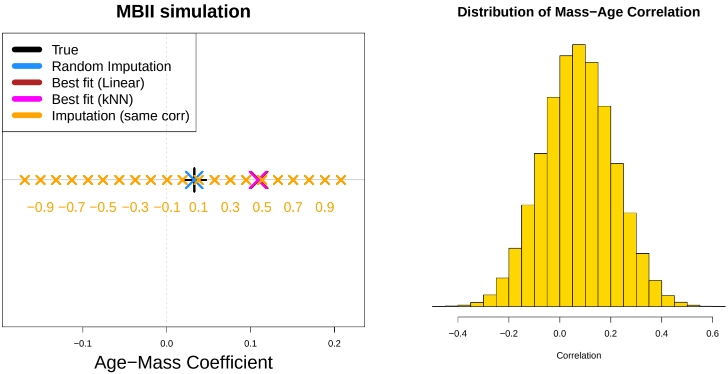
Sensitivity analysis: a partial solution

- Here is a simple method to roughly investigate the effect–we assume a single number for all correlations and evaluate how it influences our final result.

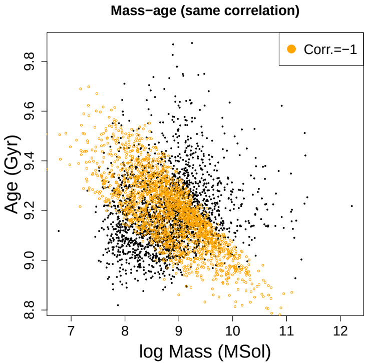


Sensitivity analysis: a partial solution

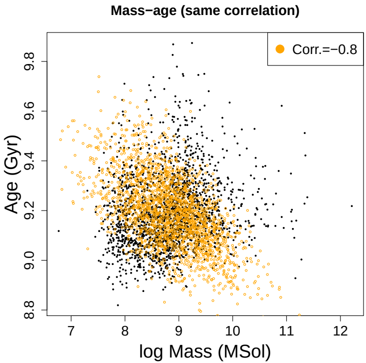
- Here is a simple method to roughly investigate the effect–we assume a single number for all correlations and evaluate how it influences our final result.



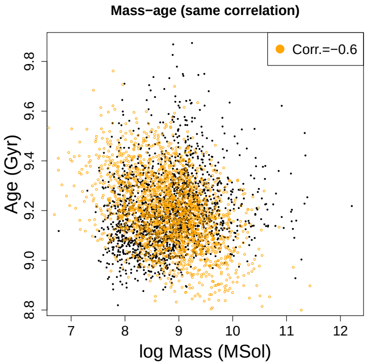
Sensitivity analysis: a graphical illustration



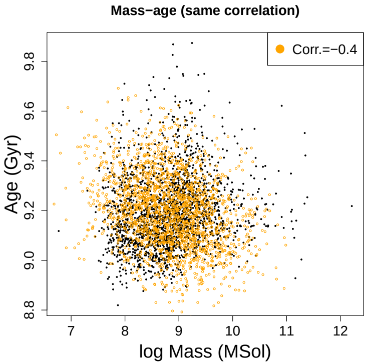
Sensitivity analysis: a graphical illustration



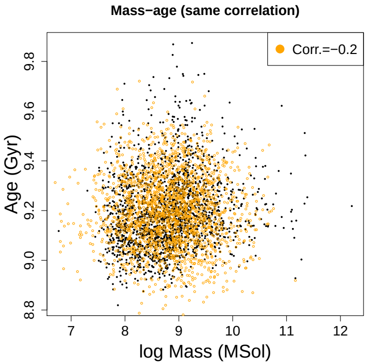
Sensitivity analysis: a graphical illustration



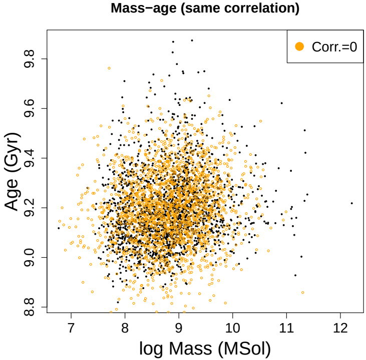
Sensitivity analysis: a graphical illustration



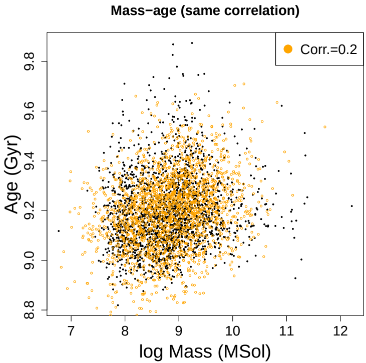
Sensitivity analysis: a graphical illustration



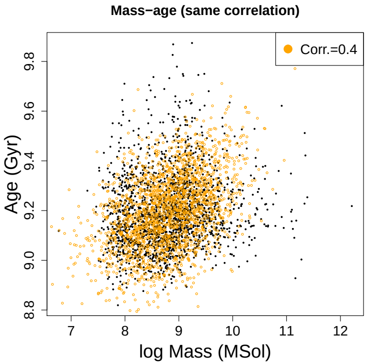
Sensitivity analysis: a graphical illustration



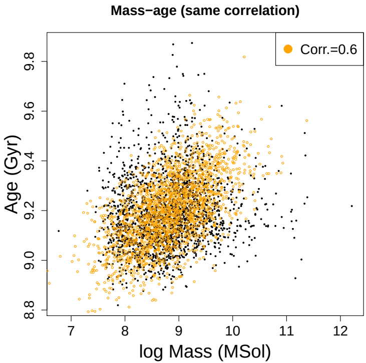
Sensitivity analysis: a graphical illustration



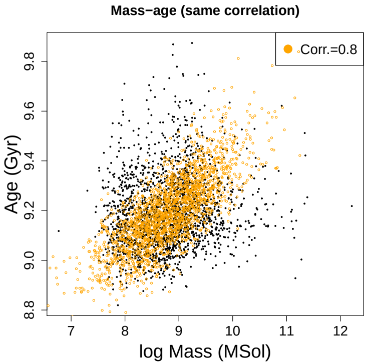
Sensitivity analysis: a graphical illustration



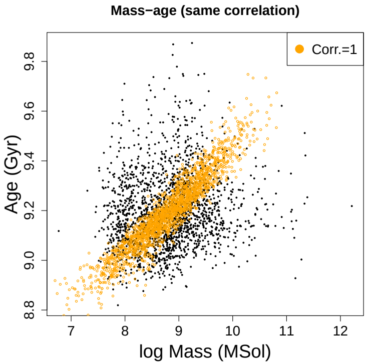
Sensitivity analysis: a graphical illustration



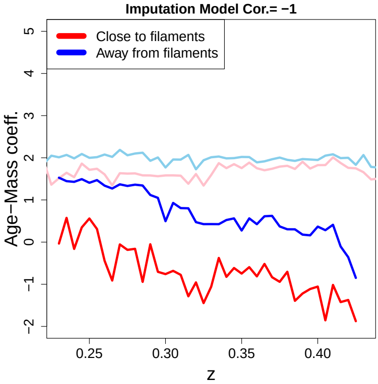
Sensitivity analysis: a graphical illustration



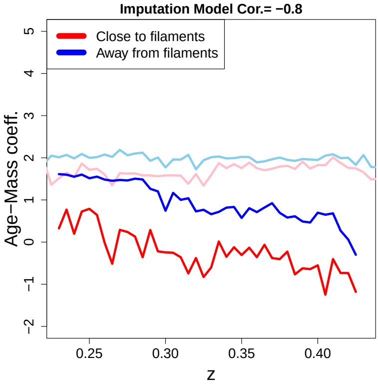
Sensitivity analysis: a graphical illustration



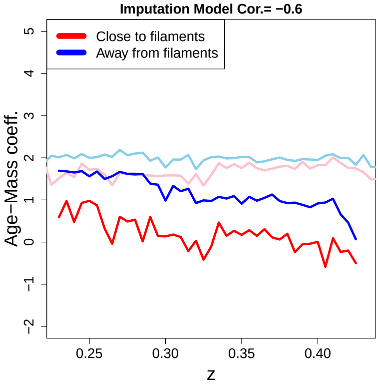
Sensitivity analysis: the SDSS data



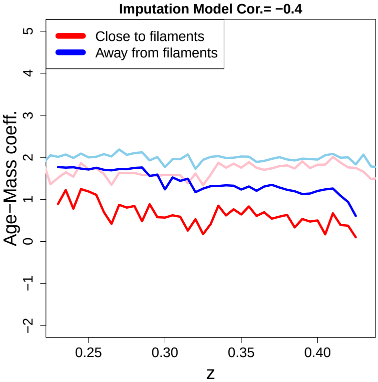
Sensitivity analysis: the SDSS data



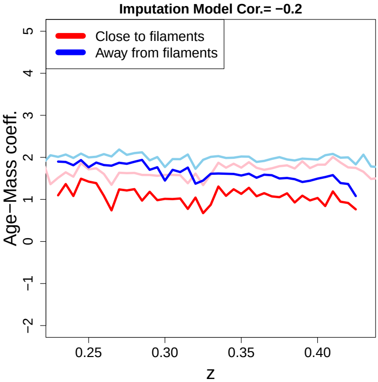
Sensitivity analysis: the SDSS data



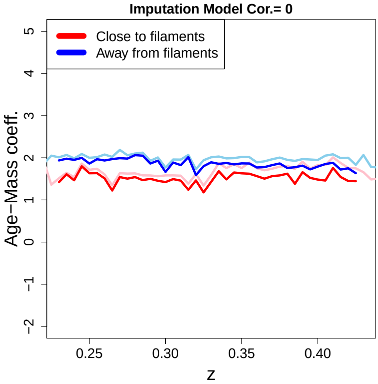
Sensitivity analysis: the SDSS data



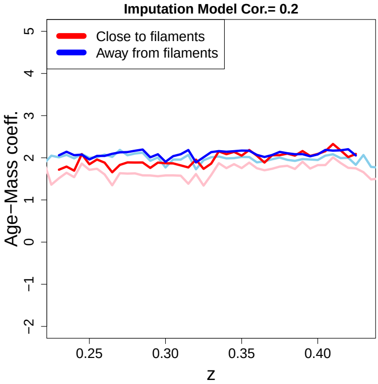
Sensitivity analysis: the SDSS data



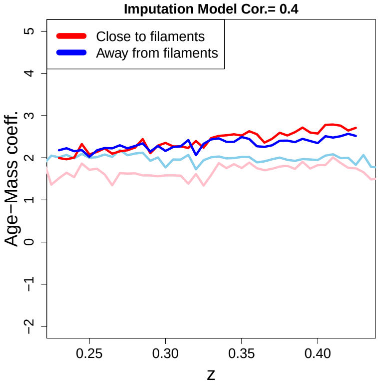
Sensitivity analysis: the SDSS data



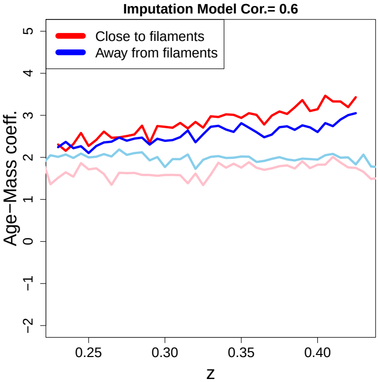
Sensitivity analysis: the SDSS data



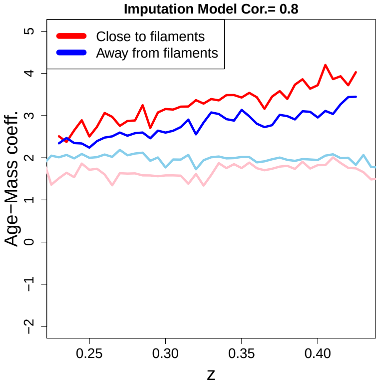
Sensitivity analysis: the SDSS data



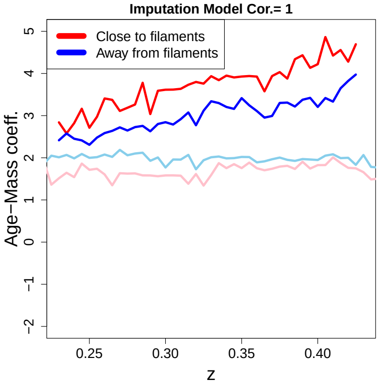
Sensitivity analysis: the SDSS data



Sensitivity analysis: the SDSS data



Sensitivity analysis: the SDSS data



Comments: sensitivity analysis

- As can be seen from the analysis on SDSS, the problem is very severe!
- We saw that the effect (from filaments) may reverse the direction if we incorrectly specify the correlation.

Comments: sensitivity analysis

- As can be seen from the analysis on SDSS, the problem is very severe!
- We saw that the effect (from filaments) may reverse the direction if we incorrectly specify the correlation.
- And the error due to the imputation is way higher than the estimation errors, which means that we should not ignore this effect.

Comments: sensitivity analysis

- As can be seen from the analysis on SDSS, the problem is very severe!
- We saw that the effect (from filaments) may reverse the direction if we incorrectly specify the correlation.
- And the error due to the imputation is way higher than the estimation errors, which means that we should not ignore this effect.
- Note: here we assume that the correlation is the same across different galaxies, but in simulation, we know that they are not the same.

Challenges of random imputation - 1

- A seemly simple solution to the above issue is that when releasing a value-added data, we also release the correlation between any pair of inferred variables.

Challenges of random imputation - 1

- A seemly simple solution to the above issue is that when releasing a value-added data, we also release the correlation between any pair of inferred variables.
- However, this would increase the number of variables a lot–if we have k inferred variables, we would have $\binom{k}{2}$ correlations.

Challenges of random imputation - 1

- A seemly simple solution to the above issue is that when releasing a value-added data, we also release the correlation between any pair of inferred variables.
- However, this would increase the number of variables a lot–if we have k inferred variables, we would have $\binom{k}{2}$ correlations.
- Moreover, this idea works only if *the normal distribution assumption is correct!*

Challenges of random imputation - 1

- A seemly simple solution to the above issue is that when releasing a value-added data, we also release the correlation between any pair of inferred variables.
- However, this would increase the number of variables a lot–if we have k inferred variables, we would have $\binom{k}{2}$ correlations.
- Moreover, this idea works only if *the normal distribution assumption is correct!*
- The normal distribution may not be correct in practice, so even if we have all correlations, our estimate may still be inaccurate.

Challenges of random imputation - 2

- Another approach to this problem is to include another set of inferred variables that are randomly drawn from conditional density.

Challenges of random imputation - 2

- Another approach to this problem is to include another set of inferred variables that are randomly drawn from conditional density.
- Suppose that we have k inferred variables, this would only require adding additionally k variables to the original data.

Challenges of random imputation - 2

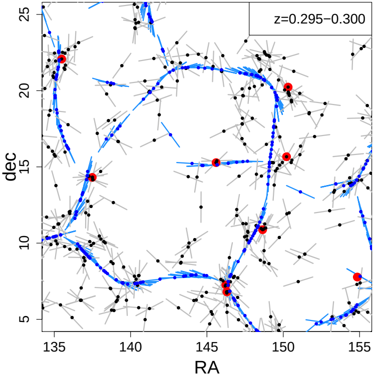
- Another approach to this problem is to include another set of inferred variables that are randomly drawn from conditional density.
- Suppose that we have k inferred variables, this would only require adding additionally k variables to the original data.
- As long as the sample size is sufficiently large, such procedure will give us a reliable estimate (Monte Carlo error will not be an issue).

Challenges of random imputation - 2

- Another approach to this problem is to include another set of inferred variables that are randomly drawn from conditional density.
- Suppose that we have k inferred variables, this would only require adding additionally k variables to the original data.
- As long as the sample size is sufficiently large, such procedure will give us a reliable estimate (Monte Carlo error will not be an issue).
- Of course, this idea relies on the assumption that the conditional density is correct, which is another strong assumption.

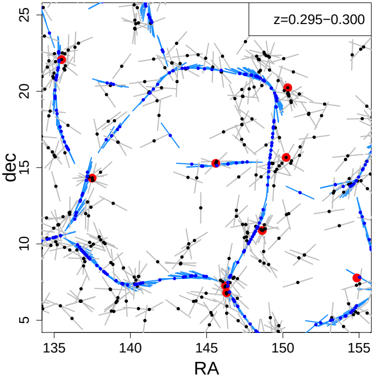
Summary

1. Statistical methods offers new exciting tools in Astronomy.



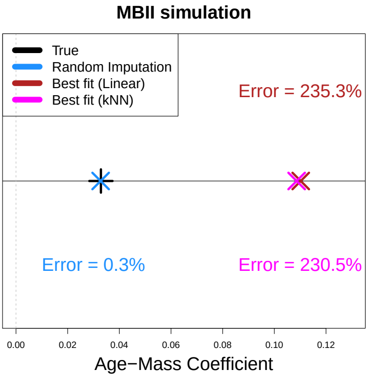
Summary

1. Statistical methods offers new exciting tools in Astronomy.
2. A good tool allows us to detect weak signals.



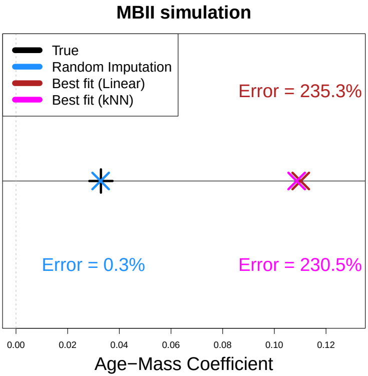
Summary

1. Statistical methods offers new exciting tools in Astronomy.
2. A good tool allows us to detect weak signals.
3. When we are using multiple derived variables, we need to be careful.



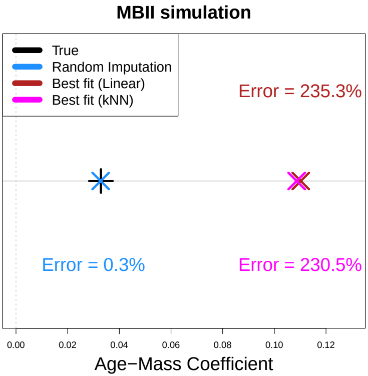
Summary

1. Statistical methods offers new exciting tools in Astronomy.
2. A good tool allows us to detect weak signals.
3. When we are using multiple derived variables, we need to be careful.
4. Using the best fitted values may result in bias in the estimation.



Summary

1. Statistical methods offers new exciting tools in Astronomy.
2. A good tool allows us to detect weak signals.
3. When we are using multiple derived variables, we need to be careful.
4. Using the best fitted values may result in bias in the estimation.
5. Random imputation offers a solution to this problem.

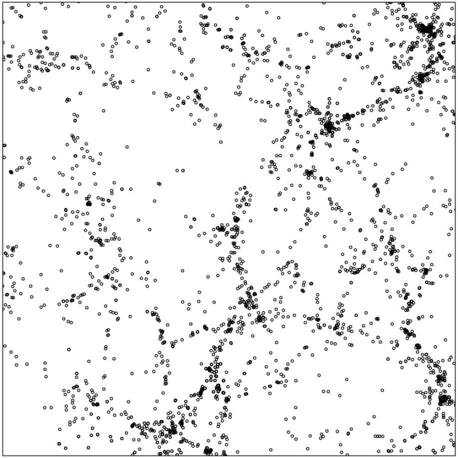


Thank you!

References

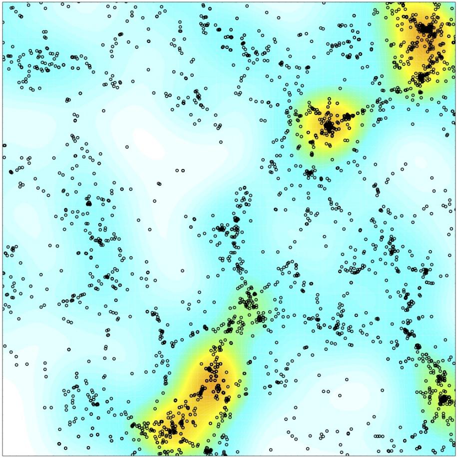
1. Joachimi, Benjamin, Marcello Cacciato, Thomas D. Kitching, Adrienne Leonard, Rachel Mandelbaum, Bjrn Malte Schfer, Cristbal Sifn et al. "Galaxy alignments: An overview." Space Science Reviews 193, no. 1-4 (2015): 1-65.
2. Tempel, E., Q. Guo, R. Kipper, and N. I. Libeskind. "The alignment of satellite galaxies and cosmic filaments: observations and simulations." Monthly Notices of the Royal Astronomical Society 450, no. 3 (2015): 2727-2738.
3. Bond, J. Richard, Lev Kofman, and Dmitry Pogosyan. "How filaments of galaxies are woven into the cosmic web." Nature 380, no. 6575 (1996): 603.
4. Chen, Yen-Chi, Shirley Ho, Peter E. Freeman, Christopher R. Genovese, and Larry Wasserman. "Cosmic Web Reconstruction through Density Ridges: Method and Algorithm." To appear in Monthly Notices of the Royal Astronomical Society..
5. Chen, Yen-Chi, et al. "Investigating Galaxy-Filament Alignments in Hydrodynamic Simulations using Density Ridges." arXiv preprint arXiv:1508.04149 (2015).
6. Chen, Yen-Chi, Christopher R. Genovese, and Larry Wasserman. "Asymptotic theory for density ridges." The Annals of Statistics 43.5 (2015): 1896-1928.
7. Conroy, Charlie, James E. Gunn, and Martin White. "The propagation of uncertainties in stellar population synthesis modeling. I. The relevance of uncertain aspects of stellar evolution and the initial mass function to the derived physical properties of galaxies." The Astrophysical Journal 699.1 (2009): 486.
8. Eberly, David. Ridges in image and data analysis. Vol. 7. Springer Science & Business Media, 1996.
9. Genovese, Christopher R., et al. "Nonparametric ridge estimation." The Annals of Statistics 42.4 (2014): 1511-1545.
10. Ozertem, Umut, and Deniz Erdogmus. "Locally defined principal curves and surfaces." The Journal of Machine Learning Research 12 (2011): 1249-1286.

1. Rawdata



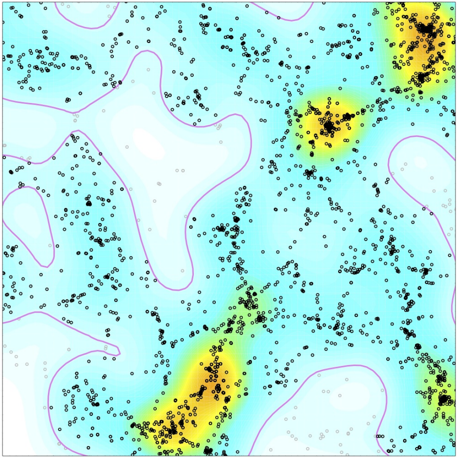
Algorithm

1. Rawdata
2. Density Reconstruction



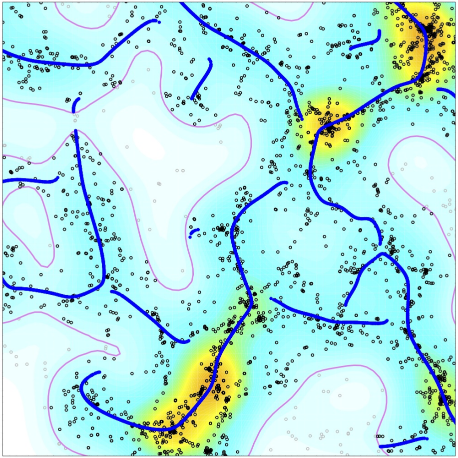
Algorithm

1. Rawdata
2. Density Reconstruction
3. Thresholding

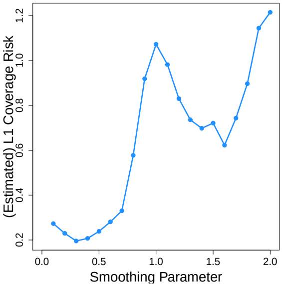


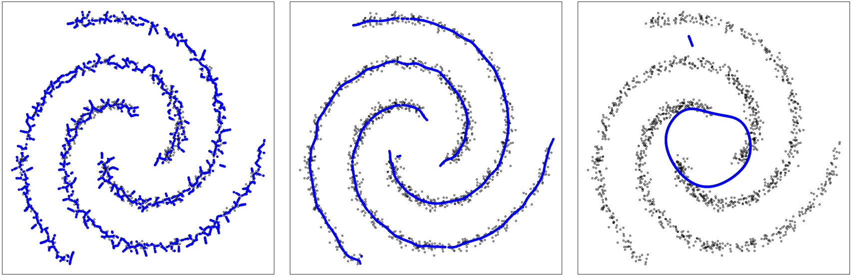
Algorithm

1. Rawdata
2. Density Reconstruction
3. Thresholding
4. Ridge Recovery

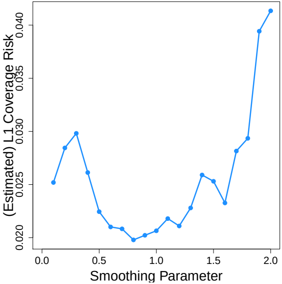


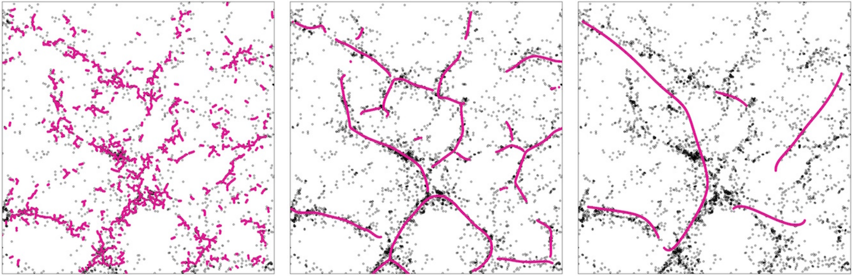
Bandwidth Selection



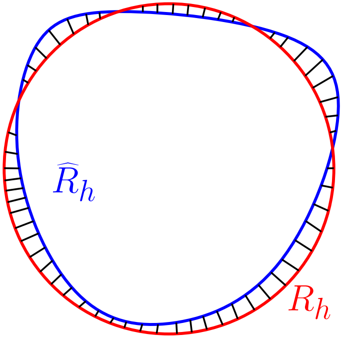


Bandwidth Selection





Bandwidth Selection



L_1 distance are like the area of the shady regions.
We estimate this distance by data splitting or the bootstrap.
Reference: **Chen** et al. 'Optimal Ridge Detection using Coverage Risk' (NIPS 2015).

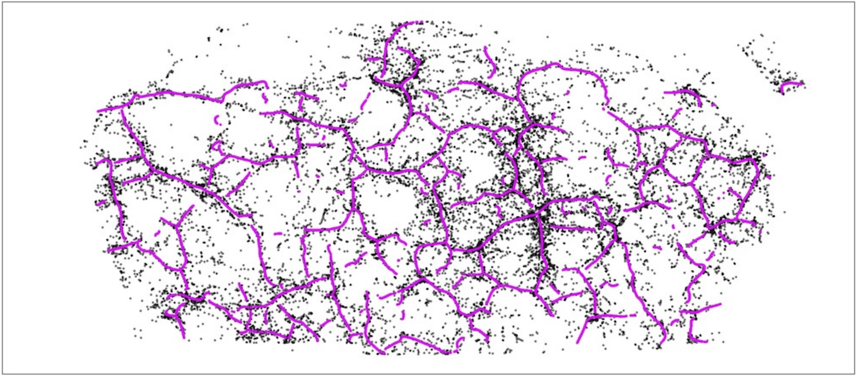
General Ridges

We can generalize ridges to higher dimensions. Pick
$V_r(x) = [v_{r+1}(x), \cdots, v_d(x)]$.
We define

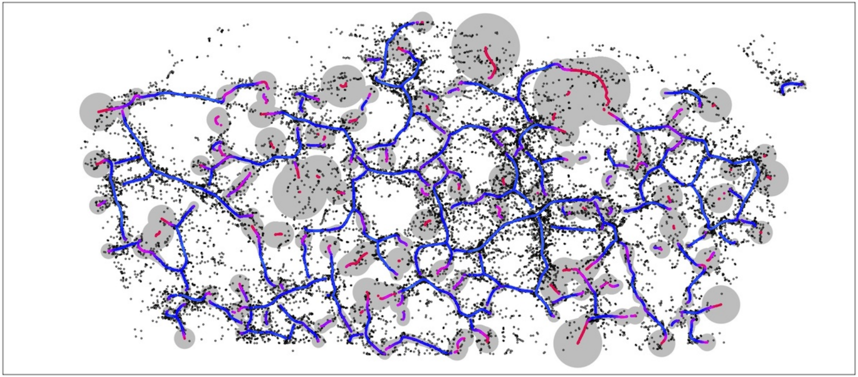
$$r\text{-Ridge}(p) = \{x : V_r(x)V_r(x)^T \nabla p(x) = 0, \lambda_{r+1}(x) < 0\}.$$

$V_r(x)$ is a $d \times (d-r)$ matrix. There are $d - r$ constraints.
By Implicit Function Theorem, r-ridges are r-manifolds.
In Astronomy, $r = 2$ can be used to model 'Cosmic Sheets (Walls)'.
$r = 0$ coincides with the definition of local modes.

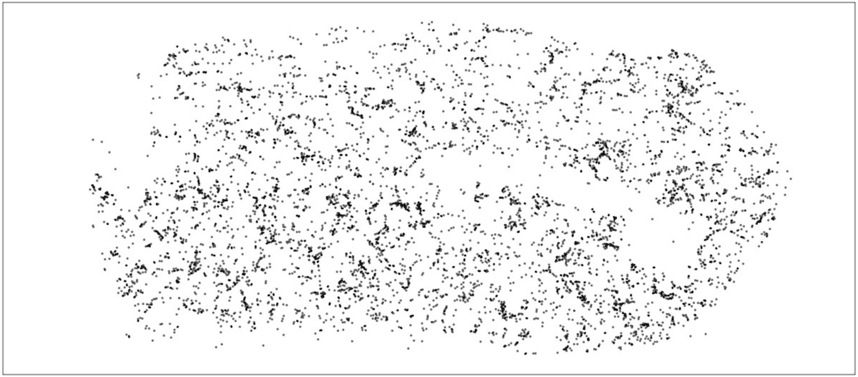
Density Ridges on the SDSS data



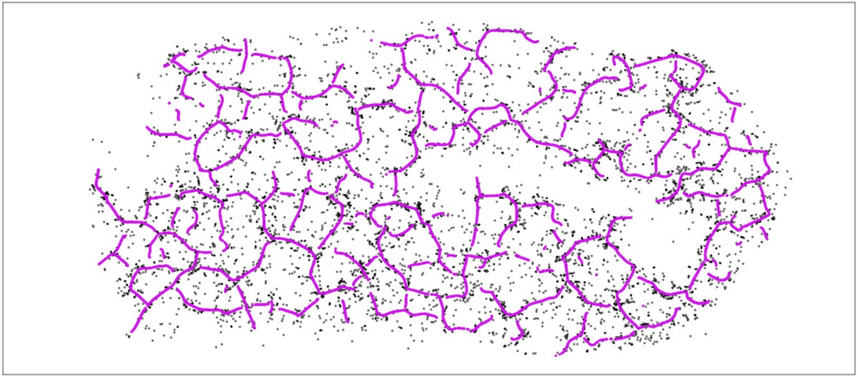
Density Ridges on the SDSS data



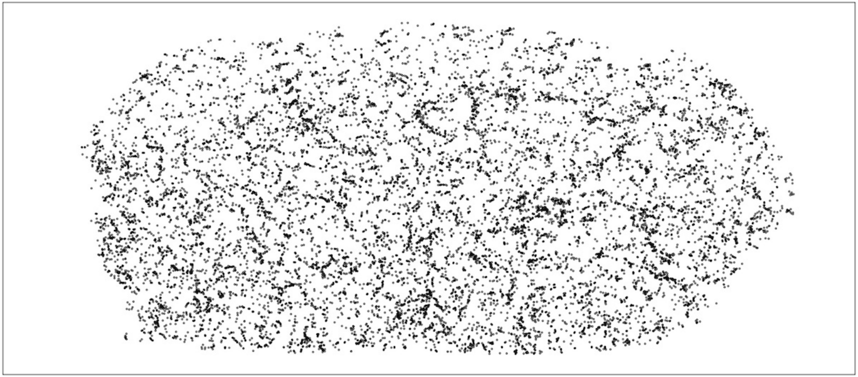
Curse of Number Density

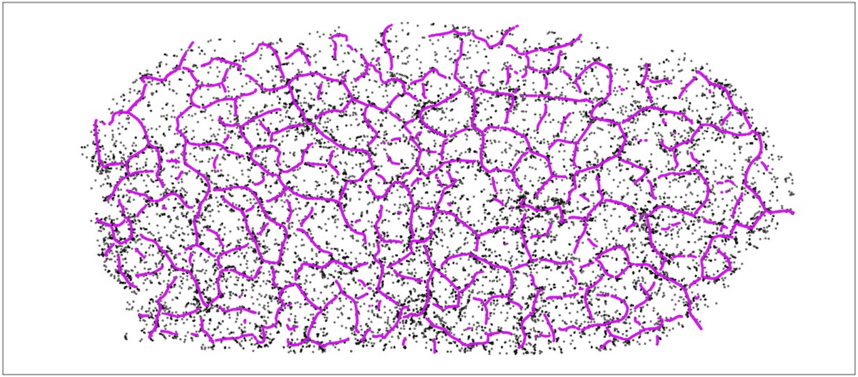


Curse of Number Density



Curse of Number Density

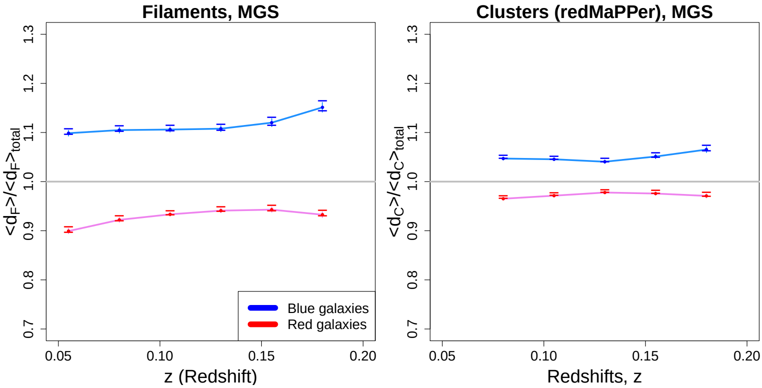




SDSS: Red and Blue Galaxies

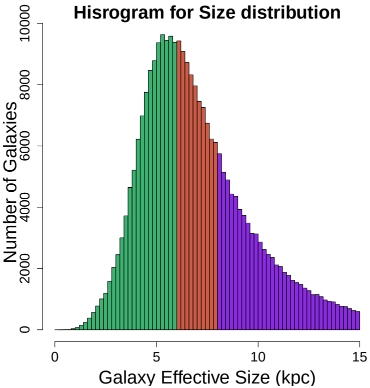
- Redshift range: $0.05 < z < 0.20$ (main sample galaxy).
- Color cut: $(g - r) = 0.8$.

SDSS: Red and Blue Galaxies

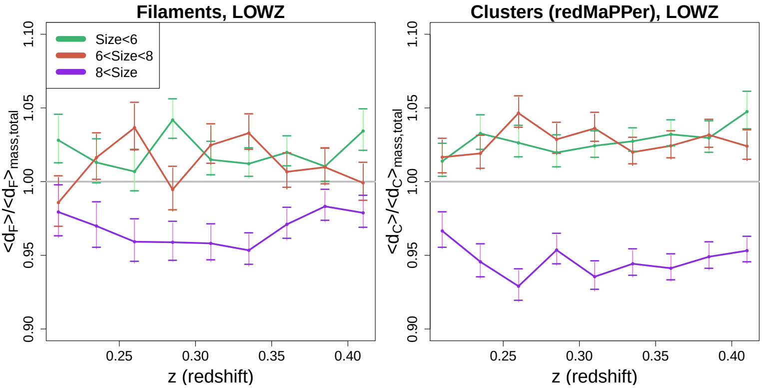


SDSS: Size for Galaxies

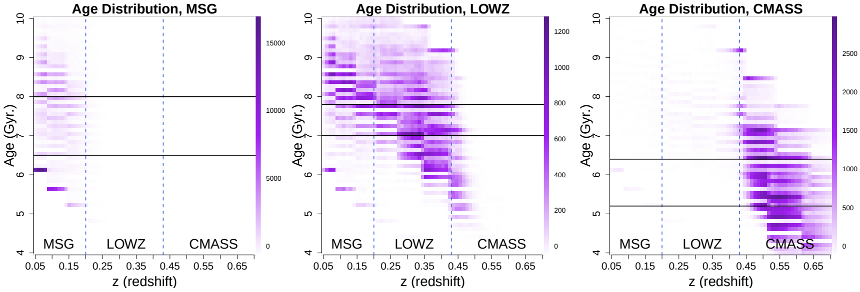
1. Size: Effective Radii.
2. Data: LOWZ ($0.20 < z < 0.43$)
3. Partitioning galaxies into three groups according to their size.



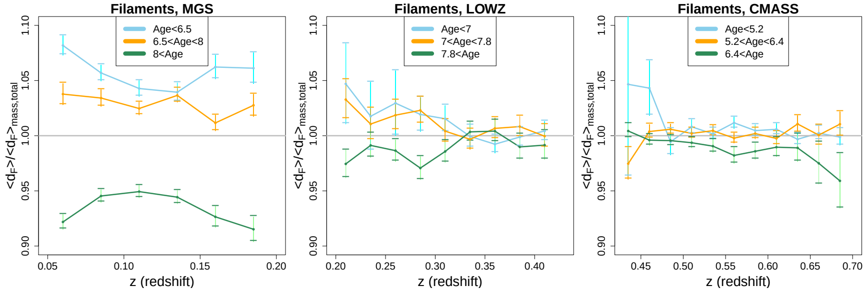
SDSS: Size for Galaxies



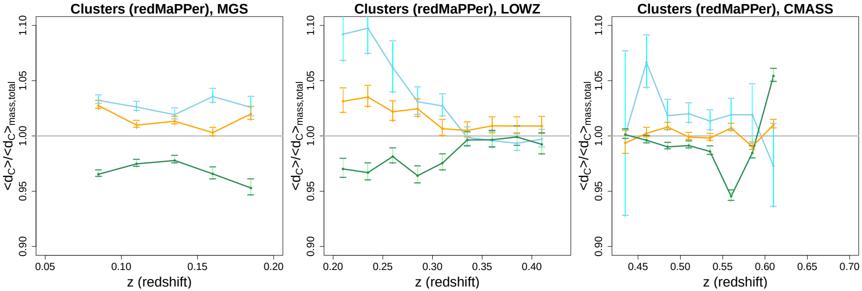
Age for Galaxies



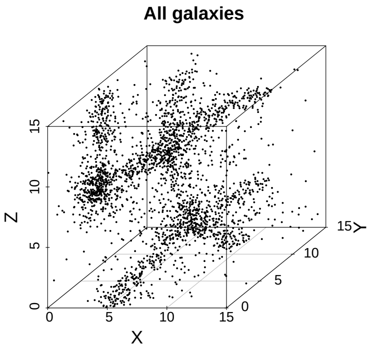
Age for Galaxies



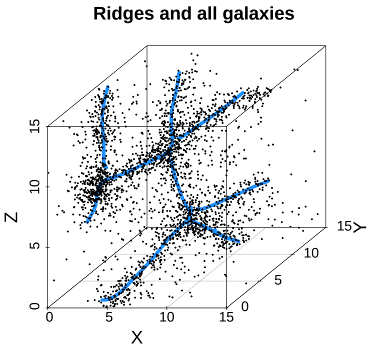
Age for Galaxies



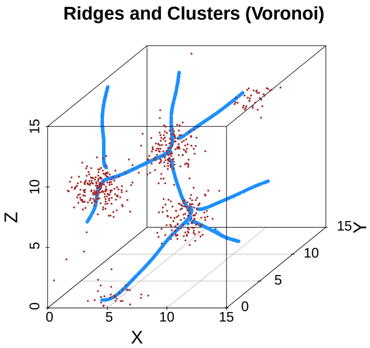
Comparison: Voronoi Model



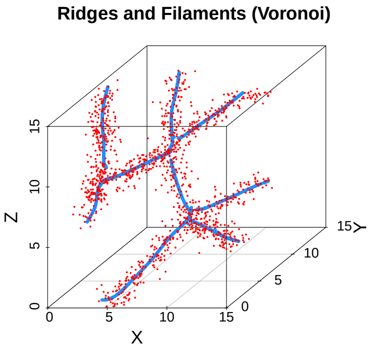
Comparison: Voronoi Model



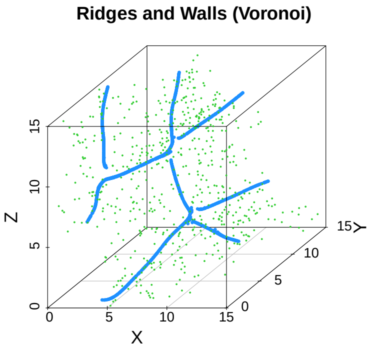
Comparison: Voronoi Model



Comparison: Voronoi Model



Comparison: Voronoi Model



Comparison: Voronoi Model