UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE CIENCIAS FÍSICAS
DEPARTAMENTO DE ASTROFÍSICA Y CIENCIAS DE LA ATMÓSFERA

TESIS DOCTORAL

New techniques for an optimal cosmological exploitation of galaxy surveys

Nuevas técnicas para explotar cosmológicamente cartografiados de galaxias de forma óptima

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Julio Jonás Chaves Montero

DIRECTORES

Carlos Hernández Monteagudo
Raúl Esteban Angulo de la Fuente

Madrid, 2018

New techniques for an optimal cosmological exploitation of galaxy surveys

Nuevas técnicas para explotar cosmológicamente cartografiados de galaxias de forma óptima

Una memoria presentada por

Julio Jonás Chaves Montero

y dirigida por

Dr. Carlos Hernández Monteagudo
Dr. Raúl Esteban Angulo de la Fuente

para aspirar al grado de Doctor en Astrofísica

DEPARTAMENTO DE ASTROFÍSICA Y CC. DE LA ATMÓSFERA
FACULTAD DE FÍSICAS
UNIVERSIDAD COMPLUTENSE DE MADRID
Madrid, julio de 2017

New techniques for an optimal cosmological exploitation of galaxy surveys

Nuevas técnicas para explotar cosmológicamente cartografiados de galaxias de forma óptima

Julio Jonás Chaves Montero

Universidad Complutense de Madrid
Departamento de Astrofísica y Ciencias de la Atmósfera

Centro de Estudios de Física del Cosmos de Aragón
Departamento de Cosmología

"Al otro, a Borges, es a quien le ocurren las cosas. Yo camino por Buenos Aires y me demoro, acaso ya mecánicamente, para mirar el arco de un zaguán y la puerta cancel; de Borges tengo noticias por el correo y veo su nombre en una terna de profesores o en un diccionario biográfico. Me gustan los relojes de arena, los mapas, la tipografía del siglo XVII, las etimologías, el sabor del café y la prosa de Stevenson; el otro comparte esas preferencias, pero de un modo vanidoso que las convierte en atributos de un actor. Sería exagerado afirmar que nuestra relación es hostil; yo vivo, yo me dejo vivir para que Borges pueda tramar su literatura y esa literatura me justifica. Nada me cuesta confesar que ha logrado ciertas páginas válidas, pero esas páginas no me pueden salvar, quizá porque lo bueno ya no es de nadie, ni siquiera del otro, sino del lenguaje o la tradición. Por lo demás, yo estoy destinado a perderme, definitivamente, y solo algún instante de mí podrá sobrevivir en el otro. Poco a poco voy cediéndole todo, aunque me consta su perversa costumbre de falsear y magnificar.

Spinoza entendió que todas las cosas quieren perseverar en su ser; la piedra eternamente quiere ser piedra y el tigre un tigre. Yo he de quedar en Borges, no en mí (si es que alguien soy), pero me reconozco menos en sus libros que en muchos otros o que en el laborioso rasgueo de una guitarra. Hace años yo traté de librarme de él y pasé de las mitologías del arrabal a los juegos con el tiempo y con lo infinito, pero esos juegos son de Borges ahora y tendré que idear otras cosas. Así mi vida es una fuga y todo lo pierdo y todo es del olvido, o del otro.

No sé cuál de los dos escribe esta página."

—Jorge Luis Borges, *Borges y yo*

"La libertad, Sancho, es uno de los más preciosos dones que a los hombres dieron los cielos; con ella no pueden igualarse los tesoros que encierra la tierra ni el mar encubre."

—Miguel de Cervantes, *Don Quijote de la Mancha*

"You pigs, you. You goof like pigs, is all. You got the most in you, and you use the least. You hear me, you? Got a million in you and spend pennies. Got a genius in you and think crazies. Got a heart in you and feel empties. All a you. Every you ...". He was jeered. He continued with the hysterical passion of the possessed. "Take a war to make you spend. Take a jam to make you think. Take a challenge to make you great. Rest of the time you sit around lazy, you. Pigs, you! All right, God damn you! I challenge you, me. Die or live and be great. Bow yourselves to Christ gone or come and find me, Gully Foyle, and I make you men. I make you great. I give you the stars."

—Alfred Bester, *The Stars My Destination*

"Alea iacta est."

—Julius Caesar, crossing the Rubicon river with his legions

A mis padres, mi tío y mis abuelos,
sois la causa de que haya llegado hasta aquí.

A ti, Virginia.

Agradecimientos

"Instruido por impacientes maestros, el pobre oye que es este el mejor de los mundos, y que la gotera del techo de su cuarto fue prevista por Dios en persona. Verdaderamente, le es difícil dudar de este mundo. Bañado en sudor, se curva el hombre construyendo la casa en que no ha de vivir..."

—Bertolt Brecht, *Loa de la duda*

Como todo el que se ha embarcado en la aventura que es realizar una tesis doctoral, pronto entendí que el desarrollo y la consecución de la misma se convertirían en el eje central de mi vida. Y que encontrarse por casualidad con tu jefe un sábado por la noche en tu despacho no sólo ocurre en las películas. Por todo ello, es lógico que tras estos cuatro años de arduo trabajo, me haya convertido poco a poco en un científico. Para ilustrar cómo fue este viaje, lo resumiré en tres etapas. Comencé con el clásico "*Sólo sé que no sé nada*", después, al ir profundizando en algunos temas, me di cuenta de que mi ignorancia era mayor aún de lo que pensaba. Por último, descubrí que lo importante es tener una visión global del problema, y que la finalidad de una tesis es poner unos granitos de arena en la gran playa que es el conocimiento. Sólo esperas que brillen fugazmente antes de que los entierre el viento.

Los responsables de que mi tesis llegase a buen puerto han sido mis dos directores, Carlos Hernández Monteagudo y Raúl Angulo. Les agradezco profundamente que me brindasen la oportunidad de comenzarla y todas sus ideas, sugerencias, enriquecedores debates y pacientes explicaciones. Sin ellos, este trabajo no se hubiera podido llevar a cabo. Y no sólo eso, también me han formado como científico y como persona. Quiero igualmente reconocer la labor del personal del Centro de Estudios de Física del Cosmos de Aragón (CEFCA), la institución donde he desarrollado la tesis. Gracias a todos por vuestros comentarios y discusiones en nuestras reuniones semanales, y por la suculenta cerveza que Javi, Toni y Carliños nos dan a catar en las"pelis con cervezas". Una mención especia se merece Silvia Bonoli, con la que he trabajado duramente desde prácticamente el inicio de mi tesis. Y cómo no, al resto de estudiantes (y Guillaume): Luis, Gonzalo, Rafa, Siddhartha, Daniele y David, con

los que he compartido largas horas de trabajo y de fiesta. También quiero alabar la labor de Mariano Moles, fundador de CEFCA, y gracias al cual muchos hemos podido dedicarnos a lo que nos apasiona. Por último, quiero valorar el apoyo económico de la Fundación Bancaria Ibercaja, que ha contribuido a que pudiese elaborar mi tesis en CEFCA.

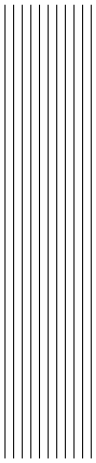
Durante el verano de 2016 completé una estancia en Leiden bajo la supervisión de Joop Schaye. Quiero agradecerle su guía y el acogerme durante dos meses en su grupo, fue de gran ayuda. Además quiero dar las gracias a todos los doctorandos en Astrofísica de la Universidad de Leiden por el "borrel" de los viernes, por los partidos de fútbol, y porque en poco menos de un mes me hicieron sentir como si llevase allí media vida.

A lo largo de nuestra trayectoria como estudiantes, nos encontramos con profesionales que ya sea por su ilusión, dedicación o los ánimos que recibimos de ellos, nos dejan un especial recuerdo. Así, quiero mencionar a Antonio Molano Romero, mi profesor de matemáticas en el instituto. Él que enseñó, como a tantos otros, a disfrutar resolviendo problemas matemáticos (y a medir una pizarra en borradores cuadrados). Asimismo me gustaría recordar a Fernando Atrio Barandela, que dirigió mi trabajo final de máster y me ayudó a dar mis primeros pasos en la investigación.

Dejando a un lado el ámbito profesional, mi principal y más importante agradecimiento es a mis padres, mi tío y mis abuelos. Siempre me habéis apoyado, creído en lo que hago, interesado por ello y valorado sin importar el resultado. Y me lo habéis demostrado desde pequeño, ya sea en la época en la que vivía en Cáceres, en las escapadas que hago a mi tierra, en las que hacéis para verme o al hablar por teléfono a diario. No puedo imaginar cómo hubiese sido yo sin vuestra influencia. Asimismo, agradezco también al resto de mi familia vuestros ánimos y los buenos ratos que pasamos juntos. Igualmente quiero agradecerte Virginia este fantástico año y todos los que nos quedan. Y darte las gracias por acompañarme en la aventura americana, seguro que nos irá muy bien. Y cómo no, no quiero olvidarme de mis amigos y reconocer su contribución a conformar lo que soy. Me es imposible nombrar a todos, pero al menos destacar a José, Luis, Samu, Paz, Curro, Carlos, Javi, Durán, Juancar, Trencitas... Hemos pasado grandes momentos juntos y lo seguiremos haciendo.

Por último quiero indicar una parte muy importante de mi vida, la música. Como decía el gran Rosendo Mercado, "Es sólo una canción y me siento mejor". Durante las al menos dos mil horas que he estado desarrollando mi tesis, prácticamente la totalidad de ellas las he pasado escuchando música. Sólo por referirme algunos grupos, me gustaría destacar a Extremoduro, Leño (y Rosendo), Loquillo, Dire Straits, Pink Floyd, Barricada, AC/DC, La Polla Records, Los Suaves, Gorillaz, Kyuss, Led Zepellin, Héroes del Silencio, Metallica, Triana, Narco, Mike Oldfield, Koma, Hora Zulú, Soziedad Alkoholika, La Frontera, System of a Down, Blue Öyster Cult, Mano Negra (y Manu Chao), Korn, Tool, Platero y Tú, The Doors, Ilegales, Mastodon y Creedence Clearwater Revival por la inspiración y hacer más llevaderas la infinidad de horas de trabajo.

Teruel, Abril de 2017



Contents

Contents xi

List of Figures xv

List of Tables xvii

Resumen de la tesis xix

Summary of the thesis xxiii

1 Introduction 1
- 1.1 Emergence of the ΛCDM model . . . 1
- 1.2 Current state of ΛCDM . . . 8
- 1.3 Open problems of ΛCDM . . . 9
  - 1.3.1 The challenges of interpreting galaxy surveys . . . 10
- 1.4 Structure of the thesis . . . 12
  - 1.4.1 Relation between galaxies and DM . . . 12
  - 1.4.2 Constraining cosmology with spectro-photometric surveys . . . 13
  - 1.4.3 Identification and redshift estimation of high-z AGN . . . 13

2 Relation between galaxies and DM 15
- Resumen en español . . . 15
- Abstract . . . 18
- 2.1 Introduction . . . 18
- 2.2 Numerical Simulations . . . 21
  - 2.2.1 The EAGLE suite . . . 21
  - 2.2.2 Catalogues and mergers trees . . . 22
  - 2.2.3 The EAGLE and DMO crossmatch . . . 23
- 2.3 Subhalo abundance matching . . . 24
  - 2.3.1 SHAM flavours . . . 24
  - 2.3.2 Implementation . . . 28

2.4 Results . . . 30
2.4.1 Correlation between $M_{\rm star}$ and V_i . . . 31
2.4.2 The properties of SHAM galaxies . . . 32
2.4.3 Galaxy clustering . . . 35
2.5 Testing the assumptions underlying SHAM . . . 42
2.5.1 The relation between $M_{\rm star}$ and V_i is independent of z . . . 42
2.5.2 Baryonic physics does not affect SHAM proxies . . . 42
2.5.3 Baryonic physics does not affect the position of subhaloes . . . 45
2.5.4 For a given $V_{\rm relax}$, $M_{\rm star}$ does not depend on environment . . . 45
2.6 Conclusions . . . 48
Appendix A: Resolution . . . 50
Appendix B: Correlation function calculation . . . 52

3 Effect of redshift errors on the galaxy clustering and the BAO 55
3.1 Introduction . . . 55
3.2 Numerical Methods . . . 56
3.2.1 Numerical Simulations . . . 56
3.2.2 Power spectrum & covariance measurements . . . 57
3.2.3 Redshift uncertainties . . . 59
3.3 Clustering with photometric redshift errors . . . 60
3.3.1 The power spectrum monopole and quadrupole . . . 60
3.3.2 The variance of the monopole and quadrupole . . . 63
3.3.3 Signal-to-noise ratio . . . 67
3.4 Effect of photo-z errors on the BAO . . . 69
3.4.1 The shape of the BAO signal . . . 70
3.4.2 Cosmological information on the BAO scale . . . 72
3.4.3 Analytical estimation of the uncertainty in α . . . 74
3.4.4 The scale-dependence of cosmological information . . . 75
3.5 Extracting information from the BAO . . . 77
3.5.1 Modelling the power spectrum monopole . . . 78
3.5.2 Parameter Likelihood Calculation . . . 79
3.5.3 Extracting the BAO scale from the simulated catalogues . . . 80
3.5.4 Constraints in cosmological parameters . . . 88
3.5.5 Effect of the PDF of photometric redshift errors . . . 89
3.6 Forecasts for future surveys with photo-z errors . . . 91
3.7 Conclusions . . . 92
Appendix A: Expressions for the effect of photo-z errors on $P(k)$ . . . 93
Appendix B: Effect of off-diagonal terms on the covariance matrix . . . 95

4 Identification and redshift estimation of high-z AGN 97
4.1 Introduction . . . 97
4.2 ELDAR algorithm . . . 99
4.2.1 Template fitting step . . . 99
4.2.2 Spectro-photometric confirmation step . . . 101
4.3 Applying ELDAR to ALHAMBRA data . . . 105
4.3.1 The ALHAMBRA survey . . . 105

4.3.2 Effects that may reduce the redshift precision and purity . . . 107
4.3.3 Specific configuration of ELDAR for the ALHAMBRA survey 109
4.3.4 Summary of the ELDAR configuration for ALHAMBRA . . . 113
4.4 The ALHAMBRA type-I AGN catalogues . . . 114
4.4.1 Properties of the ALH2L and ALH3L catalogues . . . 117
4.4.2 Quality of the ALH2L and ALH3L catalogues . . . 122
4.5 Forecasts for narrow band surveys . . . 130
4.6 Summary and conclusions . . . 132
Appendix A: AGN examples . . . 133
Appendix B: Dependence of the results on the criteria adopted in ELDAR 137
Appendix C: Description of the ALH2L and ALH3L catalogues . . . 140

5 Summary and conclusions **145**

6 Future work **149**
6.1 Optimise SHAM for emission line galaxies . . . 149
6.2 Constraining galaxy formation models using SHAM . . . 150
6.3 New reconstruction techniques for samples with redshift errors . . . 150
6.4 Spectroscopic confirmation of AGN detected at $z > 4$ by ELDAR . . . 152
6.5 Applying ELDAR to narrow-band surveys . . . 153

References **155**

Acronym list **163**

List of Figures

1.1 Large-scale distribution of dark matter 4
1.2 Current constraints from the main cosmological probes 7

2.1 $V_{\rm circ}$ of two centrals and two satellites as a function of z 25
2.2 Relation between $M_{\rm star}$ and SHAM properties for EAGLE galaxies . . 26
2.3 Properties of the Gaussians that relate $M_{\rm star}$ and SHAM parameters . 29
2.4 Spearman coefficient between $M_{\rm star}$ and four SHAM parameters . . . 30
2.5 Distribution of host halo masses for EAGLE and SHAM galaxies . . . 33
2.6 Radial distribution of EAGLE and SHAM galaxies 35
2.7 Real-space 2PCF for EAGLE and SHAM galaxies 37
2.8 Redshift-space 2PCF for EAGLE and SHAM galaxies 38
2.9 Galaxy assembly bias for EAGLE and SHAM galaxies 40
2.10 Properties of the Gaussians that relate $M_{\rm star}$ and $V_{\rm relax}$ at different z . 41
2.11 The impact on the 2PCF of different assumptions made by SHAM . . 43
2.12 Evolution of several subhalo properties along the merger history . . . 47
2.13 Number of DM particles in subhaloes of a given stellar mass 51
2.14 Number density of satellites depending on the simulation resolution . 52

3.1 Impact of photo-z errors on P_0 and P_2 62
3.2 Impact of photo-z errors on the variance of P_0 and P_2 64
3.3 Correlation matrix of P_0 for $\sigma_z = 0$ and $0.5\,\%$ 65
3.4 Correlation matrices of P_0, P_2, and $P_0 \times P_2$ for $\sigma_z = 0$ and $0.5\,\%$ . . . 66
3.5 Ratio of the SNR of $P_0(\sigma_z)$ to $P_0^r(\sigma_z = 0)$ 69
3.6 Ratio of the SNR of $P_2(\sigma_z > 0)$ to $P_2(\sigma_z = 0)$ 70
3.7 Smoothing of the BAO wiggles as a function of σ_z 72
3.8 Degeneracy between the uncertainties in the components of $\alpha_{\rm eff}$ . . . 76
3.9 Precision of our model for the BAO wiggles 78
3.10 Distribution of $\alpha_{\rm eff}$ and k_* from MCMC analyses of the BAO 80
3.11 Uncertaities in $\alpha_{\rm eff}$ from MCMC analyses of the BAO 82
3.12 Uncertainty in $\alpha_{\rm eff}$, $\alpha_{\rm eff,\perp}$, and $\alpha_{\rm eff,\parallel}$ as a function of n and σ_z 87
3.13 Constraints in Ω_m and ω derived from P_0 as a function of n and σ_z . 88

3.14 Average shift in $\alpha_{\rm eff}$ for $\sigma_z = 0.3\,\%$ as a function of the PDF 90
3.15 Estimated FoM of Ω_m and ω for surveys with photo-z errors 91
3.16 Impact photo-z errors on the precision matrix of $P_0(k)$ 95

4.1 ALHAMBRA photometry of a spectroscopically-known type-I AGN . . 102
4.2 Mock PDZ and number of AGN emission lines detected by ELDAR . . 104
4.3 Minimum EW of detectable emission lines in ALHAMBRA 105
4.4 λ_c of AGN and galaxy emission lines as a function of z 107
4.5 Mock realisations of an ALHAMBRA source with a flat SED 109
4.6 Extragalactic template database included in LEPHARE 112
4.7 Number density of the ALH2L and ALH3L catalogues 114
4.8 Source of both the ALH2L and ALH3L catalogues at $z_{\rm phot} = 1.94$. . . 115
4.9 Source of both the ALH2L and ALH3L catalogues at $z_{\rm phot} = 3.26$. . . 116
4.10 Source of the ALH3L catalogue at $z_{\rm phot} = 4.55$. 116
4.11 Redshifts and magnitudes of the ALH2L and ALH3L catalogues 118
4.12 Templates and magnitudes of the ALH2L and ALH3L catalogues . . . 119
4.13 Colour-colour diagrams for the ALH2L catalogue 120
4.14 Colour-colour diagrams for the ALH3L catalogue 121
4.15 Redshifts and magnitudes of the GAL-S, AGN-S and AGN-X samples . 123
4.16 Redshift precision for the AGN-S sample using ELDAR 125
4.17 Redshift precision for the AGN-X sample using ELDAR 126
4.18 Galaxy classified as type-I AGN by ELDAR 128
4.19 Completeness of the ALH2L and ALH3L catalogues 129
4.20 Estimated completeness for surveys with narrow-bands 131
4.21 Example of a low-z type-I AGN 133
4.22 Low-z type-I AGN with only two emission lines 134
4.23 Example of a high-z type-I AGN 135
4.24 High-z type-I AGN with only two emission lines 135
4.25 Example of a redshift outlier 136
4.26 Only type-I AGN best-fitted by a stellar template 136

6.1 Reconstruction of the BAO peak for samples with redshift errors . . . 151
6.2 SED of a type-I AGN detected at $z = 4.25$ by ELDAR 152
6.3 Quasar spectrum convolved with different filter systems 153

List of Tables

2.1 EAGLE/DMO cosmological and numerical parameters 22
2.2 Central and satellite EAGLE galaxies as a function of $M_{\rm star}$ 24
2.3 Functions that provide $M_{\rm star}$ for each SHAM implementation 28
2.4 Satellite fraction for EAGLE and SHAM galaxies 32
2.5 Number of satellites for EAGLE and SHAM galaxies 34
2.6 Effect of the stripping and star formation after infall for satellites . . 44

3.1 Degeneracy between $\alpha_{\rm eff,\parallel}$ and $\alpha_{\rm eff,\perp}$ as a function of n, σ_z, and b . . . 77
3.2 Result of the MCMC analysis of the BAO for COLA samples 81
3.3 Result of the MCMC analysis of the BAO for MXXL samples 83
3.4 Result of the MCMC analysis of the BAO as a function of n and σ_z . 85
3.5 Cosmological constraints from BAO analyses 89

4.1 Emission lines employed to confirm type-I AGN in ALHAMBRA . . . 106
4.2 Extragalactic templates included in LEPHARE 110
4.3 Equivalence between ALHAMBRA and SDSS bands 117
4.4 Redshift precision, completeness, and purity using ELDAR 124
4.5 Redshift precision for different AGN/quasar catalogues 127
4.6 Performance of ELDAR as a function of the PDZ cut-off 137
4.7 Performance of ELDAR as a function of the Ly α criterion 138
4.8 Performance of ELDAR as a function of $\sigma_{\rm line}$ 138
4.9 Performance of ELDAR as a function of $z_{\rm min}$ 139
4.10 Performance of ELDAR as a function of the magnitude limit 139
4.11 Byte-by-byte description of the ALH2L and ALH3L catalogues. . . . 142

Resumen de la tesis

"We weep for the blood of a bird, but not for the blood of a fish. Blessed are those with a voice. If the dolls could speak, no doubt they'd scream:"

"I didn't want to become human."

—Motoko Kusanagi, *Ghost in the Shell Innocence*

El desarrollo del modelo cosmológico actual, que denominamos Lambda Cold Dark Matter (ΛCDM), y la evolución de los experimentos con los que observamos el Universo iniciaron la actual era de la cosmología de precisión. Esto es debido a que ΛCDM realiza predicciones detalladas de diferentes observables en el universo cercano y lejano, lo que unido al progreso en la técnica, habilitó determinar detalladamente las propiedades del cosmos. Sin embargo, este modelo cuenta todavía con algunos problemas abiertos, como encontrar la naturaleza de la Materia Oscura (DM, Dark Matter) o descubrir qué produce la expansión acelerada del universo. En aras de solventarlos, decenas de observatorios astronómicos e instrumentos científicos se están construyendo por todo el mundo, y estos posibilitarán nuevos y más precisos cartografiados de galaxias.

Estos futuros cartografiados mapearán con exquisito detalle grandes volúmenes cosmológicos, lo que les permitirá, al extraer información cosmológica de las observaciones, reducir al mínimo las incertidumbres estadísticas. Así, los errores sistemáticos se convertirán en la principal fuente de inexactitud, donde estos emergen de interpretar erróneamente los datos o de desconocer de forma específica el impacto de las técnicas observacionales en ellos. Además, estas inexactitudes deben ser correctamente modelados para no introducir sesgos al extraer parámetros cosmológicos de los cartografiados y aprovecharlos de forma óptima.

El objetivo de la presente tesis es precisamente caracterizar algunos de estos sistemáticos y crear nuevas técnicas para obtener mayores réditos de los datos. En particular, i) analizaremos la conexión entre galaxias y halos de DM; ii) investigaremos el impacto de pequeños errores al medir el corrimiento al rojo (*redshift*) de galaxias en su distribución espacial y en la información cosmológica que se puede extraer de

ella; y iii) desarrollaremos una nueva metodología para identificar Núcleos Activos de Galaxias (AGN, Active Galactic Nuclei) y calcular su redshift usando datos de cartografiados fotométricos con bandas medianas y estrechas.

Comenzaremos estudiando la conexión entre galaxias y halos de DM empleando SubHalo Abundance Matching (SHAM), un modelo que los relaciona de forma biyectiva. Examinaremos su precisión, principales suposiciones y buscaremos su mejor implementación analizando dos simulaciones cosmológicas del proyecto Evolution and Assembly of GaLaxies and their Environments (EAGLE), la primera hidrodinámica y la segunda una versión de la primera sin bariones. Veremos que la conexión entre galaxias y halos de DM es compleja, y que una implementación cualquiera de SHAM no es capaz de reproducir el mismo agrupamiento de galaxias que encontramos en EAGLE. No obstante, generaremos una nueva implementación de SHAM que lo hace. Por lo tanto, esta puede ser utilizada para determinar las propiedades de los halos de DM que albergan galaxias detectadas en cartografiados, y a su vez extraer de forma precisa y sin sesgos información cosmológica de ellos. Asimismo, descubriremos por primera vez la existencia de *galaxy assembly bias* en una simulación hidrodinámica, donde este predice que la manera en que las galaxias pueblan halos de DM no sólo depende de la masa de estos. Por último, comprobaremos que nuestra implementación de SHAM aproximadamente captura este efecto.

En el segundo bloque modelaremos de forma teórica cómo afectan pequeñas inexactitudes al medir el redshift de las galaxias en su distribución, y confrontaremos nuestras predicciones con resultados obtenidos de cientos de simulaciones. Mostraremos que al realizar el promedio angular del espectro de potencias de la distribución de galaxias, los errores en el redshift reducen la contribución de los modos a lo largo de la línea de visión. Al estar estos más suprimidos que los perpendiculares debido a las Redshift Space Distortions (RSD), comprobaremos que la precisión con la que se pueden medir las Oscilaciones Acústicas Bariónicas (BAO, Baryonic Acoustic Oscillations) aumenta. Además, encontraremos que en el espacio de redshift la información cosmológica que se puede extraer de las BAO depende de las escalas del espectro de potencias empleadas para medirla. También veremos que la indeterminación al obtener el parámetro de Hubble es proporcional a la magnitud de los errores en el redshift. Utilizando todo lo descubierto, produciremos una metodología que permite extraer información cosmológica de cartografiados que miden el redshift de las galaxias con pequeñas inexactitudes. Por último, derivaremos una expresión que habilita estimar rápidamente la precisión en medir las BAO en función de las propiedades de la muestra de galaxias empleada.

En el último bloque crearemos un nuevo algoritmo que nos permitirá detectar AGN y calcular su redshift en cartografiados fotométricos que cuentan con filtros medianos o estrechos. Lo llamaremos Emission Line Detector of Astrophysical Radiators (ELDAR) y se basa en detectar de forma inequívoca líneas de emisión propias de AGN sirviéndose del espectro de baja resolución que estos cartografiados generan. Para caracterizar las propiedades de los AGN que nuestro método identifica, lo aplicaremos al cartografiado Advance Large Homogeneous Area Medium Band Redshift Astronomical (ALHAMBRA). Lo elegiremos debido a que observó $\sim 4\,\mathrm{deg}^2$ del hemisferio norte con 20 bandas contiguas de anchura FWHM $\simeq 300\,$Å. Al analizar ALHAMBRA con ELDAR hallaremos 494 AGN (408 de ellos desconocidos previamente) con una

densidad espacial de $176\,\mathrm{deg}^{-2}$, redshifts en el intervalo $1.5 < z < 5.5$, magnitudes más brillantes que F814W = 23, una completitud del 67 %, una precisión en redshift de $\sigma_{\mathrm{NMAD}} = 0.84\,\%$, y sin contaminación de galaxias para objetos a $z > 2$.

Como conclusión, en la presente tesis solventaremos dos desafíos a los se enfrentan los cartografiados de galaxias a la hora de extraer sin sesgos información cosmológica. El primero consiste en desentrañar cómo las galaxias trazan la distribución de materia en el universo. El segundo afronta el reto que supone modelar el impacto de pequeñas inexactitudes al medir el redshift de las galaxias en su distribución. Por último, desarrollaremos ELDAR, un método que permite detectar trazadores de la distribución de materia a alto redshift, y así posibilita determinar las propiedades cosmológicas del universo lejano.

Summary of the thesis

"We weep for the blood of a bird, but not for the blood of a fish. Blessed are those with a voice. If the dolls could speak, no doubt they'd scream:"

"I didn't want to become human."

—Motoko Kusanagi, *Ghost in the Shell Innocence*

The emergence of the current cosmological model, which we know as Lambda Cold Dark Matter (ΛCDM), and the sophistication of the experiments with which we observe the Universe enabled the actual era of precision cosmology. This is because the development of technology together with the accurate predictions of multiple cosmological observables by ΛCDM opened the possibility of determining the properties of the Universe in detail. Nonetheless, there are still some challenges in this model, e.g. to unveil the nature of Dark Matter (DM) and to explain the cause behind the accelerated expansion of the universe. To enlighten these issues, tens of observatories and scientific instruments are built all over the world, and they will produce newer and preciser galaxy surveys.

These future surveys will sample large volumes of the universe with unprecedented precision. The cosmological information extracted from them will no longer be dominated by statistical errors, and thus systematic errors will become the main source of uncertainty. These errors arise from an incorrect interpretation of the data and/or an imprecise modelling of the effect of observational techniques on the results. Moreover, they will have to be correctly accounted for in order to unbiasedly obtain cosmological information from future surveys and to fully exploit them.

This thesis precisely aims at characterising some of these systematic errors and at developing new techniques to better take advantage of the data. Particularly, i) we will study the connection between galaxies and DM haloes, ii) we will investigate the impact of sub-percent redshift errors on the galaxy clustering and on the cosmological information that encodes, and iii) we will generate a new pipeline to detect unobscured Active Galactic Nuclei (AGN) and to compute their redshifts in photometric surveys

with medium- and narrow-bands.

We will analyse the connection between galaxies and DM haloes by employing SubHalo Abundance Matching (SHAM), a model that bijectively relates them. We will examine its performance, main assumptions, and we will look for its best implementation using a cosmological hydrodynamical simulation from the Evolution and Assembly of GaLaxies and their Environments (EAGLE) project and its DM-only version. We will show that the relation between galaxies and DM haloes is not straightforward, and that a naive implementation of SHAM is not able to reproduce the same galaxy clustering as in EAGLE. On the other hand, we will define a new implementation of this model that it is able to do it to within statistical errors. Consequently, it can be used to precisely determine the properties of DM haloes that harbour galaxies from observations, and thus to extract unbiased cosmological information from them. In addition, we will find for the first time the presence of galaxy assembly bias – the dependence of galaxy clustering on properties of DM haloes beyond their mass – in a hydrodynamical simulation, and that our SHAM implementation approximately captures it.

In the second block we will theoretically model the effect of redshift errors on the galaxy clustering, and then we will confront our predictions with results from hundreds of simulations. We will show that when computing the angular average of the power spectrum of the density field, redshift errors reduce the contribution of the modes parallel to the line-of-sight. As these modes are more suppressed than the ones perpendicular to the line-of-sight due to RSD, this is translated into a better precision measuring the Baryonic Acoustic Oscillations (BAO). We will also discover that in redshift-space the cosmological information encoded in the BAO depends on the scales of the power spectrum that are employed to measure it. Furthermore, we will show that the precision measuring the Hubble parameter is inversely proportional to the magnitude of redshift errors. Using all these findings, we will generate a complete framework to extract cosmological parameters from the analysis of the BAO in galaxy survey with sub-percent redshift errors. Finally, we will also derive a theoretical expression that accurately forecasts the uncertainty measuring the BAO scale from galaxy samples with different properties.

In the last block we will develop a new methodology, which we name Emission Line Detector of Astrophysical Radiators (ELDAR), to detect AGN and to compute their redshifts in medium- and narrow-band photometric surveys. In order to do it, ELDAR takes advantage of the low-resolution spectra that this kind of surveys generate to unambiguously detect AGN emission lines. Then, we will characterise the properties of the AGN samples that ELDAR produces by applying it to data from the Advance Large Homogeneous Area Medium Band Redshift Astronomical (ALHAMBRA) survey. We will choose ALHAMBRA because it observed $\sim 4\,\mathrm{deg}^2$ of the northern sky with 20 contiguous Full Width Half Maximum (FWHM) $\simeq 300\,\text{Å}$ bands. After running our method on ALHAMBRA data, we will end up with a sample of 494 AGN (408 of them new sources) with a spatial number density of $176\,\mathrm{deg}^{-2}$, redshifts to within the interval $1.5 < z < 5.5$, magnitudes brighter than F814W = 23, a completeness of 67 %, a redshift precision of $\sigma_{\mathrm{NMAD}} = 0.84\,\%$, and no galaxy contamination at $z > 2$.

As a conclusion, in this thesis we will overcome two challenges that galaxy surveys face to unbiasedly extract cosmological information. In the first we will unravel the

way in which galaxies trace the matter distribution in the universe. In the second we will model the impact of measuring the redshift of the galaxies with noisy estimators on their distribution. Additionally, we will introduce ELDAR, a new method that detects high redshift tracers of the matter density field, and thus enables the determination of the cosmological properties of the high redshift universe.

1 Introduction

"One thing I've learned: you can know anything, it's all there, you just have to find it."

—Neil Gaiman, *Sandman*

1.1 Emergence of the ΛCDM model

Since the dawn of time, humankind has been fascinated by the night sky. The first humans surely asked themselves about what the sparkling points filling the sky were, or what the Moon was. They invoked supernatural forces to answer these questions, and every culture developed its own story about the origin of the Earth and the firmament. This kind of explanations remained the same for tens of centuries, and cosmology was only addressed by shamans and priests. We had to wait until the ancient Greeks to find critical explanations of the Universe. For the first time, they outlined theories based on observations of the sky, and their logical and mathematical interpretations. For example, Anaxagoras explained that the Sun and the stars were blazing stones, and that we do not feel the heat of the latter because they are far away from us. Moreover, Aristarchus of Samos presented the first known model that places the Sun at the centre of the Universe with the Earth revolving around it. Since them, astronomers and mathematicians worked together to elucidate what celestial bodies are and how they behave.

Physical cosmology, as commonly understood, started in the early 20th century. It began with the publication of the first modern cosmological model by Albert Einstein (Einstein 1917), which was a modified version of the field equations of General Relativity (GR) (Einstein 1916). Einstein assumed that the universe was homogeneous, filled with matter, had a positive curvature, and, in order to enable a static universe, he introduced a "cosmological constant" in his equations. Consequently, this model correctly presumed that the Universe was homogeneous; however, it kept the old hypothesis of a static and unchanging universe.

Years later, observations of "spiral nebulae" by Edwin Hubble confirmed that there were other galaxies beyond the Milky Way (Hubble 1926). Shortly after this, Georges Lemaître, inspired by dynamical cosmological models introduced by Alexander Friedmann (Friedmann 1922) and Hubble's discoveries, formulated the two hypothesis that conform the basis of our current cosmological paradigm. The first declared that the universe is expanding (Lemaître 1927), which was corroborated two years later by observations of distant galaxies (Hubble 1929), and the second that it began with a process currently known as Big Bang (BB) (Lemaître 1931).

In order to better understand an homogeneous expanding universe, let us write the Friedman-Lemaître-Robertson-Walker metric which is an exact solution of the Einstein's field equations of GR:

$$\mathrm{d}s^2 = -c^2\mathrm{d}t^2 + a(t)^2\left[\mathrm{d}r^2 + S_k(r)^2\mathrm{d}\mathbf{\Omega}^2\right], \tag{1.1}$$

where a is the cosmic scale factor, $\mathrm{d}\mathbf{\Omega}^2 = \mathrm{d}\theta^2 + \sin^2\theta\,\mathrm{d}\phi^2$ measures comoving distance, and $S_k(r)$ depends on the Gaussian curvature of the universe k as

$$S_k(r) = \begin{cases} \sqrt{k^{-1}}\sin\left(r\sqrt{k}\right), & k > 0 \text{ (closed universe)} \\ r, & k = 0 \text{ (flat universe)} \\ \sqrt{|k|^{-1}}\sinh\left(r\sqrt{|k|}\right), & k < 0 \text{ (open universe).} \end{cases}$$

Using this metric, and the Friedmann equations, we can compute the relativistic expression for the rate of expansion of an homogeneous expanding universe:

$$H^2(z) = H_0^2\left[\Omega_r(1+z)^4 + \Omega_m(1+z)^3 + \Omega_k(1+z)^2 + \Omega_\Lambda\right], \tag{1.2}$$

where $H = \dot{a}/a$ is the Hubble parameter; H_0 is the Hubble constant; $z = 1/(1+a)$ is the cosmological redshift; and Ω_i are the cosmological energy densities of radiation and relativistic particles ($i = r$), matter ($i = m$), curvature ($i = k$), and cosmological constant ($i = \Lambda$), where their total sum is normalised to unity $\Omega_r + \Omega_m + \Omega_k + \Omega_\Lambda = 1$.

Due to the confirmation of the expansion of the Universe, the cosmological constant was no longer needed, and it was rejected. In the 30s, Einstein and de Sitter adopted as fiducial cosmological model an expanding, homogeneous, isotropic, spatially flat, and matter-dominated ($\Omega_m \simeq 1$) universe (Einstein & de Sitter 1932), Einstein de Sitter (EdS) hereafter. As we will see, this model remained as the fiducial one for decades.

One of the first applications of the BB model was to explain the abundances of elements from astrophysical observations. It was soon clear that to generate light elements in early times – a process generally called Big Bang NucleoSynthesis (BBNS) – the universe must have been very hot (of the order of $10^9\,K$) and dominated by radiation (Gamow 1948; Alpher 1948). In addition, it was shown that the abundances of light elements could be used to set constraints in Ω_b (Zel'dovich 1964; Smirnov 1964; Hoyle & Tayler 1964). This is because if the value of Ω_b was close to one, nuclear reactions in the early universe would not create a detectable abundance of deuterium;

however, if it was smaller ($\Omega_b \sim 0.1$), the abundance of deuterium would be observationally detectable (Peebles 2017). In addition to his contributions to the BBNS, Gamow (1948) evolved his model for the early universe dominated by radiation to the radiation-matter equivalence, and inspired by this work, Alpher & Herman (1948) predicted a leftover radiation from the BBNS. The temperature of this radiation, which we commonly know as Cosmic Microwave Background (CMB), was expected to be very low ($5\,K$). Less than twenty years after this, Penzias & Wilson (1965) discovered the CMB and measured its temperature to be $3.5 \pm 1.0\,K$.

In the 70s, minimal estimates of the abundance of deuterium found that $0.05 < \Omega_b < 0.1$ (Geiss & Reeves 1972; Gott et al. 1974), which allowed to unambiguously certify that "Ω_b cannot exceed 0.2" (Gott et al. 1974). To maintain the condition of $\Omega_m = 1$ of the EdS universe, another form of matter that does not take part in the BBNS was needed.

The first evidence of what we know as DM came back to observations of clusters of galaxies in the 30s by Fritz Zwicky. He showed that the mass in stars was too low to explain the mass necessary to maintain clusters of galaxies in dynamical equilibrium, and thus most of their mass should not emit light (Zwicky 1933). After that, studying stars in the disk of the Andromeda galaxy, Horace Babcock showed that their rotational speed was still rising at 20 kpc from the centre of the galaxy (Babcock 1939), whereas it was expected to be decreasing if the light traces the mass distribution. Consequently, they estimated that the outer parts of Andromeda were dominated by non luminous matter, which was confirmed by 21 cm observations (van de Hulst et al. 1957) and corroborated by much detailed spectroscopic observations (Rubin & Ford 1970).

The EdS model predicted that the universe was homogeneous; nevertheless, it was clear from observations that the Universe is very clumpy on small scales, where we can see collapsed structures such as clusters of galaxies, galaxy groups, galaxies, stars, etc. The theoretical study of the evolution of perturbations in the EdS model was iniciated by Lifshitz (1946). He investigated the development of linearised inhomogeneities in an expanding universe, and, years later, Silk (1968) continued these investigations and considered that galaxies arise from primordial fluctuations of small amplitude. Then, Zel'dovich outlined his theory of the linear growth of the density and velocity fields (Zel'dovich 1970; Sunyaev & Zeldovich 1970), a.k.a "Zel'dovich approximation". He explained that collapsed structures emerge from primeval Gaussian density fluctuations, that, due to the action of gravity in an expanding universe, become non-Gaussian. In the Zel'dovich approximation, an ellipsoidal overdensity first collapses along one axis, forming a sheet; then along a second axis, forming a filament; and finally in the direction of the last axis, forming collapsed objects called haloes. This was corroborated theoretically (Peebles & Yu 1970), applied to the growth of clusters of galaxies (Gunn & Gott 1972), and confirmed with N-body simulations (Press & Schechter 1974). In Fig. 1.1 we show the large-scale distribution of structures in a modern N-body simulation, the Millenium Simulation (Springel et al. 2005). On large scales the universe is homogeneous and isotropic. On smaller scales, space is filled with filaments separated by voids, which resembles a foam-like structure usually referred to as "the cosmic web". On even smaller scales, we find DM haloes and subhaloes.

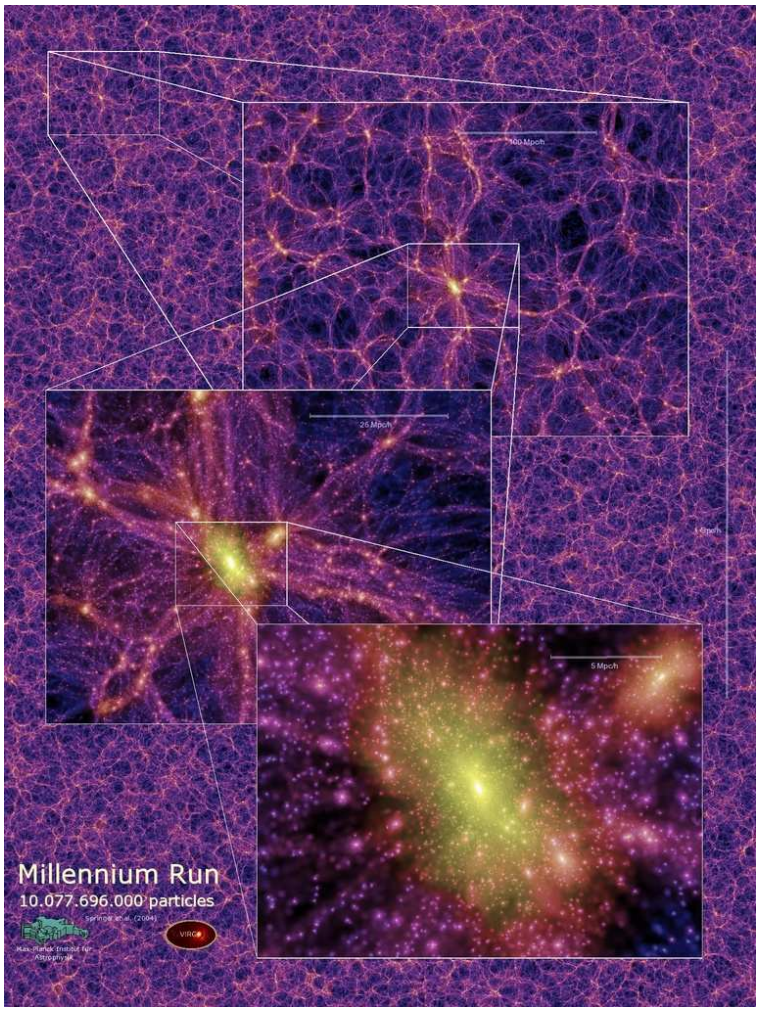


Figure 1.1: The complex distribution of DM from a N-body simulation (Springel et al. 2005). On large-scales we can appreciate that the DM density field is smooth, on intermediate-scales the emergence of voids and filaments, and on small scales the distribution of DM haloes, which may harbour galaxies.

Observational evidences for the existence of DM, which suggested that it was more abundant than baryonic matter (normal matter), and the theory of evolution of perturbations led to a new model of galaxy formation (White & Rees 1978). In this model i) the places where galaxies form and merge were determined by pure gravitational processes, and ii) the galaxy properties were given by baryonic processes. This new model assumed the presence of large amounts of DM that accounted for more than the 80 % of the energy density of the Universe. Thus the distribution of DM largely determined where galaxies form and evolve. Morevoer, it predicted a hierarchical growth of structure, i.e. the small-scale virialized systems merged together to form larger ones. Nevertheless, the DM was thought to be low-mass stars and the origin of the primeval fluctuations was not clear.

In the beginning of the 80s, there was no way to reconcile the clumpy distribution of matter at low redshift (low-z) with the smooth CMB observed at high redshift (high-z) (the upper limits for the anisotropies in the CMB were $\delta T/T < 10^{-4}$, Uson & Wilkinson 1982). In order to solve this issue, Peebles (1982) proposed that baryonic matter was subdominant with respect to a gas of nonbaryonic, weakly interacting, and massive particles. This assumption solved the problem of an homogeneous CMB and a clumpy low-z Universe because the DM perturbations could grow before the decoupling of baryonic matter and radiation, and thus they had enough time to generate the structures that we detect at low-z. This new scenario is usually referred to as the Cold Dark Matter (CDM) cosmological model. It inherits all the assumptions of the EdS model but it divides matter into baryonic matter (Ω_b) and CDM ($\Omega_{\rm dm}$).

During the first years of the 80s a new model for the early universe was also introduced: the cosmic inflation (Guth 1981; Linde 1982; Albrecht & Steinhardt 1982). This model proposed that the universe underwent an epoch of exponential growth during the first second after the BB, driven by a scalar field that slowly rolled from a local to the absolute minimum. The inflationary theory explained i) the origin the perturbations that generate the large-scale structure of the universe, ii) the horizon problem, i.e. the homogeneity and isotropy of the CMB on large scales, iii) the flatness of the universe at high-z ($\Omega_k = 0$), and the absence of magnetic monopoles. It illustrated that the primeval perturbations were quantum fluctuations that, during the inflationary phase, became classical and were converted into Gaussian perturbations of the energy density field (Starobinsky 1982; Hawking 1982; Guth & Pi 1982; Bardeen et al. 1983). Moreover, these perturbations eventually grew to form the actual large-scale structure of the universe, which was predicted to be dominated by DM (Peebles 1982; Blumenthal et al. 1984).

The next step to constrain the new CDM model was to determine $\Omega_{\rm dm}$ and Ω_b. As galaxies trace the distribution of DM, the community started to study the dynamic of the galaxy distribution to constrain cosmology. However, it was shortly noticed that the way in which galaxies trace DM was not straightforward (Kaiser 1984; Davis et al. 1985; Bardeen et al. 1986), which is still an open problem of cosmology that will be addressed in Chapter 2. Nevertheless, massive clusters could be used to constraint cosmology because they should contain the universal mix of DM and baryons. Comparing the fraction of baryons measured from galaxy clusters and from BBNS, it was found that $\Omega_{\rm dm} < 0.3$ (White et al. 1993), which was the first solid determination of

$\Omega_{\rm dm} + \Omega_b = \Omega_m < 1$. Therefore, the old assumption of $\Omega_m = 1$ inherited from the EdS model started to fail.

In the 90s, X-ray observations nearby galaxy clusters (Myers et al. 1997), the measurement of the redshift-magnitude relation from standard candles, and precise observations of the CMB power spectrum definitively rejected $\Omega_m = 1$. Type Ia supernovae are cataclysmic explosions caused by white dwarfs that gradually accrete material from another star, surpass the Chandrasekhar limit – at this point the electron degeneracy pressure is unable to prevent the collapse of the star – raising the temperature of their cores and starting the fusion of carbon, and then fiercely explode. Using them, which are supposed to be standard candles, i.e. to have exactly the same luminosity independently of their properties and redshift, two independent groups corroborated that $\Omega_m < 1$. To explain the energy content of the universe (remember that $\Sigma_i \Omega_i = 1$), they brought back the Einstein's old idea of a cosmological constant $\Omega_\Lambda > 0$ (Riess et al. 1998; Schmidt et al. 1998; Perlmutter et al. 1999). Nonetheless, this was not the only discovery pointing towards $\Omega_m < 1$.

The angular power spectrum of the CMB was an old prediction of the linear evolution of cosmological perturbations (Sachs & Wolfe 1967), and for the first time it was measured during these years, e.g. The Tenerife Experiment, *COsmic Background Explorer* (COBE), Australia Telescope Compact Array (ATCA), the Toco experiment, Balloon Observations Of Millimetric Extragalactic Radiation and Geophysics (BOOMERanG), and Millimeter-wave Anisotropy Experiment Imaging Array (MAXIMA). The results were consistent with the CDM model (Peebles 1982; Bond & Efstathiou 1984), and much smaller than earlier predictions for models in which $\Omega_b > \Omega_{\rm dm}$ (Sachs & Wolfe 1967; Peebles & Yu 1970). In addition, the CMB was predicted to have oscillatory features, called BAO. They were generated by perturbations of the photon-baryon hot plasma in the epoch prior to recombination, and then imprinted in the CMB (Peebles & Yu 1970; Sunyaev & Zeldovich 1970). Once the observations of the CMB became more precise, the first peak of the BAO was detected. Its amplitude and position, which encode cosmological information, corroborated that the value of Ω_m was smaller than one, specifically $0.25 < \Omega_m < 0.50$ (Balbi et al. 2000).

The inflationary theory, the discovery of $\Omega_\Lambda > 0$, which implies that the Universe is undergoing an accelerated phase of expansion, and the existence of CDM as the main gravitating component in the universe constitute the ΛCDM model. Impressively, ΛCDM is able to simultaneously explain: i) the measurements of the Hubble parameter at $z < 0.01$, ii) the galaxy dynamics at $z < 0.1$, iii) the supernovae magnitude redshift relation at $z < 1$, and iv) the angular power spectrum of the CMB at $z > 1000$. Arguably, the most remarkable aspect of this model is that it produces testable predictions for a large variate of observations. This and the technical advances in the experiments with which we observe the universe have fuelled the start of a new era in cosmology, where the free parameters of the ΛCDM model are measured with exquisite precision and its assumptions tested thoroughly.

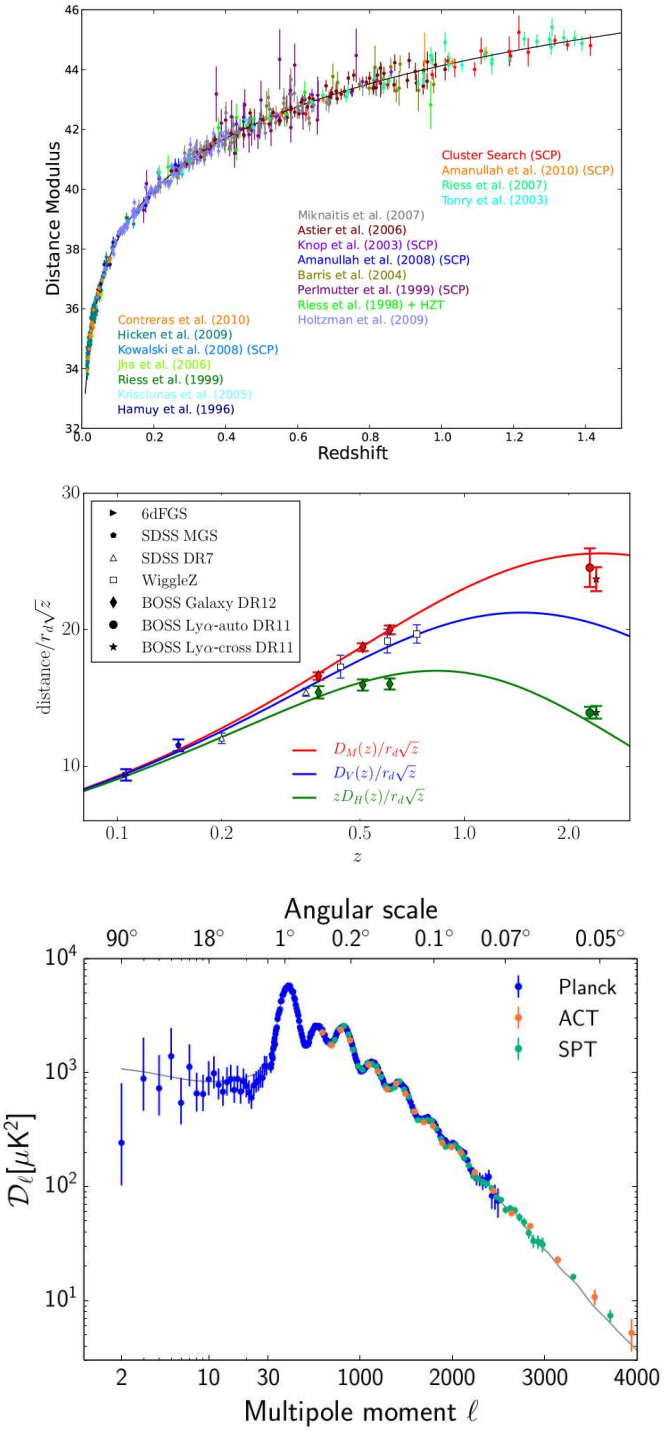


Figure 1.2: Relation between distance modulus and redshift for type Ia supernovae at $z < 1.5$ (Suzuki et al. 2012) (top panel), cosmological results extracted from BAO detected in the galaxy clustering at $z < 0.7$ and the Ly α forest of quasars at $z \sim 2$ (Alam et al. 2016) (middle panel), and the power spectrum of the CMB anisotropies at $z \sim 1100$ (Planck Collaboration et al. 2016b) (bottom panel). The ΛCDM model, which is indicated by solid lines, is in agreement with all the observations.

1.2 Current state of ΛCDM

The simplest form of the ΛCDM model assumes that i) the law of gravity is given by GR, ii) the cosmological constant is responsible of the accelerated expansion of the universe, iii) the DM is formed by high mass (cold) and low self-interacting (collisionless) particles, and iv) an inflationary phase in the early universe during which the seeds of the large-scale structure were generated. In its minimal expression, it features 6 free parameters: the physical baryon density $\Omega_b\, h^2$, the physical DM density $\Omega_c\, h^2$, the age of the universe t_0, the scalar spectral index n_s, the dimensionless curvature power spectrum A_s, and the reionization optical depth τ. In addition, it presumes flatness and adiabatic perturbations. During the last few decades, multiple cosmological observations have constrained the numerical values of these parameters with an ever increasing precision and accuracy. The main probes of ΛCDM are CMB, galaxy clustering, Weak gravitational Lensing (WL), and type Ia supernovae. In what follows we will shortly describe each one.

As we mentioned before, in the early 90s the CMB anisotropies were first measured. Since then, they have become one of the most powerful probes of cosmology and the physics of the early universe, and several spacecrafts, such as *Wilkinson* Microwave Anisotropy Probe (WMAP) and *Planck*, and ground base experiments, e.g. Atacama Cosmology Telescope (ACT) and South Pole Telescope (SPT), have pursued their characterisation. The CMB anisotropies are divided in two main types: temperature and polarization anisotropies. The power spectra of both types can be computed with high precision using linear perturbation theory, which is very precise describing the early universe, and thus they can be used to set robust cosmological constraints.

A simple deep image of the sky provides the two dimensional galaxy clustering; however, the redshifts of the galaxies need to be measured in order to compute its three dimensional counterpart. Measured redshifts encode a combination of the velocity due to the expansion of the universe, i.e. the Hubble flow, and the peculiar velocity caused by the gravitational attraction by large-scale structure. As a consequence, peculiar velocities alter the otherwise isotropic galaxy clustering. The three dimensional galaxy clustering allows to measure BAO and RSD. The BAO were first detected by Two-degree-Field Galaxy Redshift Survey (2dFGRS) (Percival et al. 2001; Cole et al. 2005) and Sloan Digital Sky Survey (SDSS) (Eisenstein et al. 2005), and they provide information about the Hubble parameter and the angular diameter distance, which is important because measurements of the Hubble parameter at different redshifts precisely constrain the accelerated expansion of the universe (Blake & Glazebrook 2003; Hu & Haiman 2003; Linder 2003; Seo & Eisenstein 2003). The RSD constrain the growth history of the universe (Percival & White 2009) and enable to test GR on large scales (Raccanelli et al. 2013).

Another cosmological probe that can be addressed by galaxy surveys is WL. WL by large-scale structure has been discussed for a long time (Gunn 1967; Miralda-Escude 1991); however, it was not measured until the beginning of this century (Kaiser et al. 2000; Wittman et al. 2000; Van Waerbeke et al. 2000; Bacon et al. 2000). It relies on the measurement of the correlations between the shapes of galaxies, which provides information about the expansion and growth history of the universe. This is because

the photons emitted by distant galaxies, in their path towards us, are perturbed by the matter distribution, which distorts the shape of the distant galaxies. Consequently, the matter distribution can be mapped by measuring these distortions.

As we mentioned before, type Ia supernovae are assumed to be standard candles. Consequently, they inform us about the expansion history of the Universe. Moreover, as type Ia supernovae are among the brightest objects in the Universe, they may be used to constrain cosmology from low to high redshift (the most distant type Ia supernovae ever detected is at $z = 2.26$ Rodney et al. 2015).

The *Planck* best estimates for the the minimal set of free ΛCDM parameters are the following (Planck Collaboration et al. 2016a): $\Omega_b\, h^2 = 0.02225 \pm 0.00016$, $\Omega_c\, h^2 = 0.1198 \pm 0.0015$, $t_0 = 13.799 \pm 0.021\,\mathrm{Gyr}$, $n_s = 0.9645 \pm 0.0049$, $\ln(10^{10} A_\mathrm{s}) = 3.094 \pm 0.034$ at $k_0 = 0.05\,\mathrm{Mpc}^{-1}$, and $\tau = 0.079 \pm 0.017$. In Fig. 1.2 we show cosmological constraints derived from observations of type Ia supernovae (top panel), spectroscopic wide-field surveys (medium panel), and the CMB (bottom panel). It is remarkable that the ΛCDM model, indicated by solid lines, is able to precisely fit all these observations.

1.3 Open problems of ΛCDM

Despite its simplicity, success, and predicting power of ΛCDM, the model is still unsatisfactory for several reasons. Firstly, the hypothetical DM particle is yet to be directly found. Secondly, very little is known about the fields driving inflation, and we are yet to detect their signatures in form of primordial tensor modes and specific non-Gaussianities. Finally, the nature of the accelerated expansion of the universe is unknown and it could be caused by the failure of GR on large scales.

Decisive evidence for the DM particle could come from their direct detection in subterranean laboratories, or indirectly by observing their possible self annihilation or decay into standard model particles. On a parallel front, wide-field surveys are set to constrain the abundance and global properties of DM. The cosmic DM abundance can be determined from RSD and WL, the collisional cross section from merging clusters, and the DM particle mass from the abundance of satellite galaxies and the Ly-alpha forest (because structure is suppressed on scales below the free streaming length).

The inflationary period in the early universe should produce tensor modes that are in principle detectable in the polarisation of CMB photons. Such a detection could be a smoking gun of inflation and the energy scale at which occurred. Alternatively, galaxy surveys could also constrain the physics of inflation via measurements of the primordial spectral index, and of non-Gaussianities in the primordial density field. Both of these quantities are expected to be related to the properties and number of inflationary fields.

Finally, the nature of the agent causing the accelerated expansion of the universe is still not clear. Wide-field surveys offer an opportunity to decipher this mystery. The BAO peak is a cosmic standard ruler than can measure the expansion history of the universe in detail, and RSD probe the relation between density and velocities fields, and thus they test the law of gravity on cosmological scales.

It is clear from the above that galaxy surveys can potentially shed light on the ΛCDM model, potentially finding new fundamental physics and substantially improv-

ing our understanding of the universe. Ongoing and future surveys such as Dark Energy Survey (DES), Extended Baryon Oscillation Spectroscopic Survey (eBOSS), Subaru Hyper Suprime-Cam (HSC), Hobby-Eberly Telescope Dark Energy Experiment (HETDEX), Javalambre Physics of the Accelerating Universe Astrophysical Survey (J-PAS), Dark Energy Spectroscopic Instrument (DESI), WEAVE, Euclid, Subaru Prime Focus Spectrograph (PFS), Large Synoptic Survey Telescope (LSST), and Wide-Field Infrared Survey Telescope (WFIRST) will take on this opportunity.

1.3.1 The challenges of interpreting galaxy surveys

To capitalise on the opportunity offered by galaxy surveys, several challenges on different levels must be solved. Firstly, to correctly model the relation between galaxies and DM haloes. Secondly, to find how cosmological probes encode cosmological information and how to extract it in an unbiased manner. Thirdly, to find observables at different redshifts which may be used to constrain cosmology.

The goal of this thesis is to provide further knowledge that could help tackling these issues. Specifically, we will explore the connection between DM structures and the stellar mass of the hosted galaxies; the impact of redshift uncertainties on the BAO; and develop an efficient algorithm to identify AGN and to compute their redshifts, which would be used to trace the density field at high-z.

As we commented before, the two most important probes in galaxy surveys are galaxy clustering and WL. Here we will describe possible systematics affecting them.

The precision of cosmological constraints extracted from galaxy surveys relies on our theories about the propagation of sound waves in the early universe, the non-linear evolution of the matter density field, and the way in which galaxies trace DM. On the one hand, density fluctuations in the early universe are well described by linear perturbation theory and their physics is robustly and precisely tested by the CMB (Planck Collaboration et al. 2014b, 2016a). On the other hand, the non-linear evolution of the matter density field shifts the position of the BAO scale computed from linear theory (Crocce & Scoccimarro 2008; Smith et al. 2008). In addition, the connection between galaxies and DM is not straightforward, e.g. galaxy formation may also shift the BAO scale (Padmanabhan & White 2009; Tseliakhovich & Hirata 2010; Mehta et al. 2011) and van Daalen et al. (2014) showed that different prescriptions for baryonic processes in simulations modify the galaxy clustering on average a 10 % on small scales. And this is not only true for clustering analyses, Eifler et al. (2015) estimated that cosmological constraints extracted from WL in LSST and Euclid may be biased as much as $\sim 7\,\sigma$ if these processes are not taken into account. In Chapter 2 we will model the connection between galaxies and DM on non-linear scales using cosmological simulations. This will allow future galaxy surveys to extract extract unbiased cosmological information.

Galaxy surveys will also face observational systematics. They are classified into two main categories according to the strategy employed to scan the sky: photometric and spectroscopic surveys. The first take large images of the sky with a few filters, and their main advantages are that they reach deeper magnitudes for the same integration time, allow faster mapping speeds, and produce redshifts for a much greater number of astrophysical objects than spectroscopic surveys. Furthermore, photometric surveys

observe every pixel of the sky, whereas spectroscopic surveys require a preselection of the sources. On the other hand, spectroscopic surveys produce much preciser redshifts than photometric surveys, which allow them to accurately measure the three dimension galaxy clustering, and thus RSD and BAO along the line-of-sight (e.g., Rodríguez-Torres et al. 2016; Beutler et al. 2017b).

In the last years, spectro-photometric surveys has emerged. This new kind of survey takes images of the sky using multiple medium- and/or narrow-band filters, e.g. Classifying Objects by Medium-Band Observations - a spectrophotometric 17-filter survey - (COMBO-17) (Wolf et al. 2004, 2008), The Cosmic Evolution Survey (COSMOS) (Ilbert et al. 2009), ALHAMBRA (Moles et al. 2008), Survey for High-z Absorption Red and Dead Sources (SHARDS) (Pérez-González & Cava 2013), The Physics of the Accelerating Universe Survey (PAUS) (Martí et al. 2014), and J-PAS (Benítez et al. 2014). Although spectro-photometric surveys are neither as deep as photometric surveys nor as precise measuring redshifts as spectroscopic surveys, they combine the main characteristics of both to produce a low-resolution spectra for every pixel of the sky. The precision with which they measure redshifts increases with the number of contiguous bands (Benítez et al. 2009b), e.g. COSMOS achieves a redshift precision of $\sigma_z/(1+z) = 0.8\,\%$ for galaxies with $i^+ < 23$. Future narrow-band survey such as PAUS and J-PAS are expected to reach even a higher precision, which will allow them to measure BAO along the line-of-sight (Benítez et al. 2009a). In Chapter 3 we will explore the effect of sub-percent redshift errors on the galaxy clustering in general and on the BAO in particular. This will enable to extract unbiased information from BAO analyses in spectro-photometric surveys.

Several wide-field surveys have measured BAO from the clustering of galaxies at low-z. On the other hand, it is not straightforward to use galaxies to detect BAO at high-z ($z > 2$). This is because as they become fainter, they are harder to detect and their redshifts more difficult to compute. Currently, the only measurements of BAO at high-z are done by using the Ly α forest (Busca et al. 2013), which is a collection of absorption features in the spectra of high-z quasars – optically unobscured AGN with emission lines – caused by the absorption of neutral hydrogen between the quasar and us. The amount of absorption indicates the density of neutral hydrogen, and it can be used to measure the large-scale structure of the universe. Moreover, ongoing and future surveys will directly employ the three dimensional distribution of quasars to detect BAO (e.g., eBOSS is expected to reach a 1.6 % precision measuring spherically averaged BAO with them, see Dawson et al. 2016; Zhao et al. 2016).

To detect high-z AGN, photometric surveys usually employ colour-colour selection techniques (Matthews & Sandage 1963) and/or intrinsic variability studies (Schmidt et al. 2010). Spectroscopic surveys preselect AGN candidates using colour-colour diagrams, observe them, and confirm the ones that pass a visual inspection (Pâris et al. 2014, 2017). Finally, spectro-photometric surveys use the multiple colours that can be constructed with their bands, and in some cases additional data from other wavelengths (e.g., Salvato et al. 2009). In Chapter 4 we will introduce a new method to detect high-z quasars and to compute their redshifts, which will enable to obtain samples with higher redshift precision and lower contamination from galaxies and stars.

In the following section we provide further details of the structure of the thesis

and our main findings.

1.4 Structure of the thesis

This thesis is divided into three main blocks. In each one of them, we will address one of the main challenges in galaxy surveys introduced in the previous section. In Chapter 2 we will study and model the connection between galaxies and DM, in Chapter 3 we will develop a complete framework to extract cosmological information from spectro-photometric surveys, and in Chapter 4 we will introduce a new methodology to detect high-z unobscured AGN with emission lines. In Chapter 5 we will summarise our main findings and present the main conclusions of this thesis. Finally, in Chapter 6 we will outline ongoing and future work that continues the lines of investigation opened in this thesis. In the following we summarise the main findings in each of the three major chapters of this thesis.

1.4.1 Relation between galaxies and DM

We already noticed that the relation between galaxies and DM is not straightforward on small scales. The best predictions come from hydrodynamical simulations, i.e. N-body simulations that evolve together DM and baryons (see Kuhlen et al. 2012, for a review). Nevertheless, the largest ones that produce realistic results encompass volumes too small for clustering studies, e.g. the state-of-the-art EAGLE (Schaye et al. 2015) and Illustris (Vogelsberger et al. 2014) simulations evolve a comoving volume of $\sim 10^{-3}\mathrm{Gpc}^3$, whereas to robustly detect BAO a greater volumes has to be sampled (at least $1\,h^{-3}\mathrm{Gpc}^3$, see Tegmark 1997). Moreover, on scales smaller than $30\,h^{-1}\,\mathrm{Mpc}$ the cosmological results produced by current models for RSD are biased (White et al. 2015), and greater scales cannot be samples with current hydrodynamical simulations.

In Chapter 2 we will model the connection between DM haloes and galaxies using SubHalo Abundance Matching (SHAM) (Vale & Ostriker 2004; Shankar et al. 2006; Conroy et al. 2006). We will examine its performance, main assumptions, and look for its best-implementation using two cosmological simulations from the EAGLE suite (Schaye et al. 2015; Crain et al. 2015), the first with and the second without baryons. We will generate a new implementation of SHAM that produces the same galaxy clustering as the EAGLE hydrodynamical simulation. In addition, we will detect for the first time galaxy assembly bias in an hydrodynamical simulation, and we will show that our SHAM implementation approximately capture it.

The applications of our new SHAM implementation are threefold. It can be used to estimate the halo mass of galaxies from observations, to model the effect of baryonic physics on cosmological probes, and to constrain galaxy formation models by fitting its results to observations.

This chapter was published in a referred astronomy journal under reference Chaves-Montero et al. 2016, MNRAS, 460, pp. 3100-3118.

1.4.2 Constraining cosmology with spectro-photometric surveys

The impact of photo-z errors on the galaxy clustering has been explored by several authors (e.g., Seo & Eisenstein 2003; Blake & Bridle 2005; Cai et al. 2009; Benítez et al. 2009a). In configuration space, they smooth the galaxy field along the line-of-sight, and in Fourier space they reduce the amplitude of the line-of-sight modes. Nevertheless, the BAO scale can still be measured along the line-of-sight, for instance, the error in the BAO scale only doubles for galaxy samples with $\sigma_z/(1+z) = 0.3\,\%$ with respect to spectroscopic samples with the same number density (Cai et al. 2009). As spectro-photometric surveys will produce redshifts for a larger number of astronomical objects than spectroscopic surveys, they may provide even better cosmological constraints than spectroscopic surveys. However, their effect has to be correctly modelled.

In Chapter 3 we will explore the impact of sub-percent photo-z errors on the galaxy clustering in Fourier space, developing a complete methodology for the exploitation of the BAO signal and its cosmological interpretation. We will demonstrate that they modify the cosmological information encoded in the BAO scale in a scale-dependent manner. Furthermore, we will demonstrate that redshift errors reduce the contribution of power spectrum modes parallel to the line-of-sight when computing its angular average. This translates into a better precision measuring the BAO scale from samples with sub-percent errors when the contribution of the shot-noise is subdominant. We will also provide a fitting function that forecasts the precision with which the BAO scale can be measured as a function of the underlying cosmology and the number density, large-scale bias, and photo-z error of the analysed galaxy sample. Moreover, we will show that this fitting function reproduces numerical results obtained from hundreds of N-body simulations.

The most straightforward application of the presented framework will be to extract unbiased cosmological information from BAO analysis in spectro-photometric surveys. Additionally, the introduced fitting function will allow to search the galaxy sample that provides the best cosmological constraints, which is usually not the one with the smallest photo-z errors. This function could be used to optimally design future galaxy surveys too.

This chapter is under referee revision in the astronomy journal MNRAS.

1.4.3 Identification and redshift estimation of high-z AGN

AGN are very bright sources powered by the accretion of matter onto the central SuperMassive Black Hole (SMBH) of a galaxy, and some of them are so bright that can be detected at very high-z (currently, the most distant spectroscopically confirmed quasar is at $z = 7.1$, see Mortlock et al. 2011).

In Chapter 4 we will produce a new pipeline to identify unobscured AGN with emission lines and to compute their redshifts in spectro-photometric surveys. We will take advantage of the low-resolution spectra provided by these surveys to automatically detect AGN emission lines, and thus confirm them. Furthermore, we will test our new methodology, dubbed as ELDAR, with data from the ALHAMBRA sur-

vey (Moles et al. 2008; Molino et al. 2014). We will choose ALHAMBRA because it covered 8 non-overlapping fields using 20 contiguous bands. Thus, this is a good benchmark for future narrow-band surveys. We will characterise the completeness, redshift precision, and galaxy contamination in the produced samples, and we will detect hundreds of previously-unknown type-I AGN.

As a conclusion, we provide a new method to detect AGN in spectro-photometric surveys and we demonstrate its efficacy. The ultimate goal of ELDAR is to produce samples of high-z AGN to extract cosmological constraints from their clustering, which will be feasible in future survey like PAUS and J-PAS (Abramo et al. 2012).

This chapter is under referee revision in the astronomy journal MNRAS.

2 Relation between galaxies and DM

– "We're in a loop."
– "Yes, I know, there are time anomaly leaks everywhere, but we're not in one right now. Are we?"
– "No, not horizontal loop. Vertical one."
– "What do you mean?"
– "Look through my microscope. And then through my telescope. You'll see."

—Mateusz Skuntnik, *Submachine*

This chapter has been published as Chaves-Montero et al. 2016, MNRAS, 460, 3100C

Resumen en español

Conocer el modo en que las galaxias se agrupan nos permite estudiar diversas propiedades del Universo. En escalas pequeñas (desde varios kiloparsecs a unos pocos megaparsecs), la función de correlación de las galaxias (o su equivalente en el espacio de Fourier, el espectro de potencias) nos informa acerca del modo en que la materia bariónica puebla las estructuras de DM. Así, nos posibilita inferir con qué eficiencia se forman las estrellas en función del tiempo cosmológico, qué procesos regulan dicha eficiencia y cómo interactúan las galaxias con su medio. En escalas más grandes, nos ayuda a discernir la composición del Universo, su historia de expansión y el funcionamiento de la gravedad. Por último, en escalas aún mayores, nos proporciona conocimiento acerca de cómo fueron los primeros instantes del Universo (periodo inflacionario) y nos permite poner a prueba las predicciones de la Relatividad General.

Sin embargo, la extracción de esta información es una tarea compleja como ya hemos comentamos en la introducción, debido a que el modo en que las galaxias se agrupan depende de la cosmología y de los procesos que atañen su formación y

evolución. Esto es especialmente cierto en escalas pequeñas, donde las interacciones son altamente no lineales y se forman y fusionan halos de DM y galaxias. No obstante, no sólo encontramos procesos fundamentalmente gobernados por la gravedad como los anteriores, sino también otros que son debidos a la presencia de bariones, véase la formación de estrellas, las explosiones de supernovas, la emisión de energía por parte de AGN o la reducción del suplemento de gas en las galaxias como consecuencia de su interacción con el gas caliente de grupos y cúmulos. Dada la complejidad y variedad de todos ellos, es necesario emplear simulaciones para estudiarlos (para un extenso análisis del tema ver Kuhlen et al. 2012).

En la literatura, podemos encontrar principalmente dos acercamientos a esta cuestión. En el primero se evoluciona de manera conjunta la DM y la bariónica. Para ello, se resuelven de forma acoplada las ecuaciones de Poisson y Euler y se introducen diversas recetas para procesos físicos que no se pueden resolver en las escalas que emplea la simulación. Algunos ejemplos de éstos son la formación de estrellas, el enfriamiento radiativo y la inyección de energía por medio de supernovas y agujeros negros. Este tipo de simulaciones, comúnmente denominadas hidrodinámicas, reproducen de forma precisa diferentes propiedades galácticas (Vogelsberger et al. 2014; Schaye et al. 2015). Sin embargo, a día de hoy no es posible evolucionar volúmenes cosmológicos lo suficientemente grandes para estudiar el agrupamiento de galaxias es escalas lineales.

Con el objetivo de acceder a volúmenes mayores y además hacerlo de manera rápida, la segunda aproximación sólo simula el avance del espacio de fases de la DM, incluyendo las galaxias a posteriori. Esto está justificado ya que los halos de DM determinan dónde se forman y cómo evolucionan las galaxias (White & Rees 1978). Por otra parte, las relaciones entre estas estructuras no son triviales, lo que produce que los resultados de este segundo tipo de modelos sean más inciertos.

En este artículo, estudiaremos las predicciones y suposiciones de un modelo que se encuadra en este segundo acercamiento y cuyo nombre es SHAM (e.g., Vale & Ostriker 2004; Shankar et al. 2006; Conroy et al. 2006). Además, trataremos de encontrar su implementación más precisa. Para ello, utilizamos dos simulaciones pertenecientes al proyecto EAGLE. La primera la denominaremos EAGLE y es una simulación hidrodinámica que evoluciona un volumen cosmológico de $10^6\,\mathrm{Mpc}^3$. La segunda, que llamaremos Dark Matter Only version of EAGLE (DMO), es una versión de la primera sin bariones y que emplea su misma cosmología y condiciones iniciales. Ambas nos permitirán saber qué galaxias se hubiesen situado en qué halos, y de esta forma dilucidar qué propiedades de los halos y las galaxias están más interrelacionadas. Así, definiremos una nueva implementación de SHAM, que aplicándola a DMO, nos permitirá conocer con qué precisión nuestro nuevo modelo es capaz de recuperar la distribución espacial de galaxias en EAGLE.

Veremos que nuestra implementación produce una función de correlación estadísticamente compatible con la de EAGLE para galaxias con $\log_{10} M_{\mathrm{star}}[\mathrm{M}_\odot] > 10.77$. Por contra, para galaxias menos masivas y en escalas menores a 2 Mpc, encontraremos que sobreestimamos su valor en un 30 %. Detectaremos la presencia de galaxy assembly bias en EAGLE, lo cual nunca había sido confirmado en una simulación hidrodinámica, y comprobaremos que modifica la función de correlación de las galaxias en un 20 %. Además, constataremos que nuestro modelo reproduce este efecto con un error no mayor al 15 %. Finalmente, descubriremos que las pequeñas diferencias entre la

distribución de galaxias que produce SHAM y la que medimos de EAGLE aparecen debido a que algunas de las suposiciones de SHAM no se cumplen totalmente.

A modo de resumen, en este trabajo presentaremos una nueva implementación de SHAM que es capaz de reproducir el efecto de la física bariónica en la distribución de galaxias con gran precisión. Una de las posibles aplicaciones será poblar simulaciones de DM que cubren grandes volúmenes cosmológicos con galaxias para así poder modelizar el impacto de la física bariónica en distintos observables cosmológicos. Otras aplicaciones obvias son estimar las propiedades de los halos donde residen galaxias observadas en cartografiados y, si configuramos nuestro modelo con diferentes simulaciones hidrodinámicas, dilucidar qué modelos de formación de galaxias son más precisos. Para ello, sólo habría que comparar el agrupamiento de galaxias observado en cartografiados con el que produce SHAM tras calibrarlo con diferentes simulaciones hidrodinámicas.

Subhalo abundance matching and assembly bias in the EAGLE simulation

Chaves-Montero et al. 2016, MNRAS, 460, 3100C

ABSTRACT: SHAM is a widely-used method to connect galaxies with DM structures in numerical simulations. SHAM predictions agree remarkably well with observations, yet they still lack strong theoretical support. We examine the performance, implementation, and assumptions of SHAM using the EAGLE project simulations. We find that $V_{\rm relax}$, the highest value of the circular velocity attained by a subhalo while it satisfies a relaxation criterion, is the subhalo property that correlates most strongly with galaxy stellar mass ($M_{\rm star}$). Using this parameter in SHAM, we retrieve the real-space clustering of EAGLE to within our statistical uncertainties on scales greater than 2 Mpc for galaxies with $8.77 < \log_{10}(M_{\rm star}[{\rm M}_\odot]) < 10.77$. Conversely, clustering is overestimated by 30 % on scales below 2 Mpc for galaxies with $8.77 < \log_{10}(M_{\rm star}[{\rm M}_\odot]) < 9.77$ because SHAM slightly over-predicts the fraction of satellites in massive haloes compared to EAGLE. The agreement is even better in redshift-space, where the clustering is recovered to within our statistical uncertainties for all masses and separations. Additionally, we analyse the dependence of galaxy clustering on properties other than halo mass, i.e. the assembly bias. We demonstrate assembly bias alters the clustering in EAGLE by 20 % and $V_{\rm relax}$ captures its effect to within 15 %. We trace small differences in the clustering to the failure of SHAM as typically implemented, i.e. the $M_{\rm star}$ assigned to a subhalo does not depend on i) its host halo mass, ii) whether it is a central or a satellite. In EAGLE we find that these assumptions are not completely satisfied.

2.1 Introduction

The clustering of galaxies offers an excellent window to explore galaxy formation processes and the fundamental properties of our Universe. On small scales, correlation functions can inform us about the way in which galaxies populate DM haloes and thus about the efficiency of star formation and the importance of environmental effects.

On large scales, the clustering of galaxies can be used to constrain cosmological parameters and the law of gravity. On even larger scales, the observed distribution of galaxies is sensitive to the physics of inflation and relativistic effects. By using correlation functions of different orders and at distinct scales, degeneracies among several parameters can be broken, providing even tighter constrains on all the aforementioned quantities.

To extract the information encoded in the clustering of galaxies, we need accurate predictions for a given cosmological scenario and galaxy formation model. However, obtaining the correct galaxy distribution is a difficult task, especially at small scales where besides highly non-linear dynamics, gravitational collapse, mergers, dynamical friction, and tidal stripping; baryonic processes such as star formation, feedback, and ram pressure are at play. Consequently, one needs to resort to numerical simulations to obtain accurate predictions for galaxy clustering (see Kuhlen et al. 2012, for a review).

Two types of approach can be followed. The first is to simulate the joint evolution of DM and baryons by solving the Poisson and Euler equations coupled with recipes for unresolved physical processes (e.g. star and black hole formation). Although this approach currently yields the most direct predictions for the distribution of galaxies, it is computationally infeasible to simulate large cosmological volumes with adequate resolution for calculating accurately the galaxy clustering on scales on the order of 100 h^{-1} Mpc. In addition, simulations have only recently begun to produce populations of realistic galaxies (Vogelsberger et al. 2014; Schaye et al. 2015).

The second approach is to simulate only gravitational interactions and to predict the galaxy clustering a posteriori. This is justified by leading theories of galaxy formation, where DM plays the dominant role in determining the places where galaxies form and merge. Gravity-only simulations (a.k.a. DM-only simulations) are computationally less expensive and can thus follow sufficiently large volumes to enable the correct interpretation of observational surveys. This is an important advantage since, for instance, to model galaxy clustering on scales beyond 100 h^{-1} Mpc, it is necessary to perform N-body simulations of volumes in excess of 1 h^{-3} Gpc^3 (Angulo et al. 2008). The disadvantage is that the predictions for galaxy clustering are more uncertain because the relation between galaxies and DM haloes is not straightforward.

Subhalo abundance matching (e.g., Vale & Ostriker 2004; Shankar et al. 2006; Conroy et al. 2006) is a widely-used method to populate gravity-only simulations with galaxies. The original version of SHAM assumes an injective and monotonic relation between galaxies and self-bound DM structures based on a set of specified properties. SHAM usually links galaxies to DM structures using stellar mass as galaxy property and a measure of subhalo mass, such as circular velocity, as subhalo property. More recent implementations introduce stochasticity into the relation to make the model more realistic (e.g., Behroozi et al. 2010; Trujillo-Gomez et al. 2011; Reddick et al. 2013; Zentner et al. 2014). Then, SHAM places each galaxy at the centre-of-potential of its corresponding subhalo and assumes that each galaxy has the same velocity as the centre-of-mass of its linked subhalo. SHAM thus makes predictions for the clustering of galaxies, but not for any physical properties such as stellar mass, star formation rate, metallicity, etc.

SHAM predictions have been shown to agree remarkably well with observations

(e.g. Conroy et al. 2006; Guo et al. 2010; Wetzel & White 2010; Moster et al. 2010; Behroozi et al. 2010; Trujillo-Gomez et al. 2011; Watson et al. 2012; Nuza et al. 2013; Reddick et al. 2013). For instance, Conroy et al. (2006) showed that SHAM reproduces the observed galaxy clustering over a broad redshift interval ($0 < z < 5$). More recently, Reddick et al. (2013) achieved a simultaneous fit to the clustering and the conditional stellar mass function measured in SDSS. Simha & Cole (2013) even used this model to constrain cosmological parameters, finding values in good agreement with those obtained from more established methods.

Despite these successes, the comparison with simulations of galaxy formation has not been so encouraging. Weinberg et al. (2008) found that the galaxy clustering predicted by SHAM only agrees with that of a hydrodynamical simulation beyond $1\ h^{-1}\,\mathrm{Mpc}$. On smaller scales, the differences were of the order of a few. Simha et al. (2012) extended the previous study using two hydrodynamic simulations with different feedback models. They found that the clustering predicted by SHAM exceeded that of their most realistic simulation by more than a factor of 2 on scales below $0.5\ h^{-1}\,\mathrm{Mpc}$. Finally, in a direct comparison with two semi-analytic models of galaxy formation, Contreras et al. (2015) found that SHAM performs well at some galaxy number densities, but not at others.

It is therefore not clear whether SHAM is able to match the observed galaxy clustering because it makes accurate assumptions (i.e. the physical relation between subhaloes and galaxies) or because some implementations employ free parameters (e.g. a scatter between subhalo and galaxy properties or a cut-off in the fraction of satellite galaxies) that provide enough freedom to become insensitive to them. The importance of the information being decoded, added to the fact that the amount and accuracy of clustering data will increase dramatically over the next decade due to the emergence of wide-field galaxy surveys (e.g. DES, HETDEX, eBOSS, J-PAS, DESI, Euclid, and LSST), makes it crucial to critically test the assumptions underlying SHAM.

In this paper we will employ the state-of-the-art hydrodynamical simulations EAGLE (Schaye et al. 2015; Crain et al. 2015) to study the SHAM technique in detail. Our objectives are threefold, i) to seek the most accurate implementation of SHAM, ii) to directly test the underlying assumptions, and iii) to assert how accurately SHAM can predict galaxy clustering.

We will propose $V_{\rm relax}$, defined as the maximum of the circular velocity of a DM structure along its entire history while it fulfils a relaxation criterion, as the best subhalo property with which to perform SHAM. We will show that this definition captures the best qualities of previously proposed implementations while mitigating their disadvantages and reducing the number of problematic cases. As a consequence, $V_{\rm relax}$ shows the strongest correlation with the simulated stellar mass of EAGLE galaxies.

We will show that SHAM is able to reproduce the clustering properties of stellar mass selected galaxies in the EAGLE simulation (which successfully reproduces many properties of observed low-z galaxies). For the stellar mass range investigated ($10^{8.77} < M_{\rm star}[\mathrm{M}_\odot] < 10^{10.77}$), the agreement is better than 10 % on scales greater than 2 Mpc, and better than 30 % on smaller scales. The agreement is particularly good for massive galaxies and in redshift space, for which we do not find statistically significant difference between the clustering predicted by SHAM and EAGLE. This is remarkable given that we explore almost two orders of magnitude in spatial scale and four in

clustering amplitude.

Additionally, we will pay attention to the so-called "assembly bias": the dependence of the clustering of DM haloes on properties other than mass (Gao et al. 2005; Zhu et al. 2006; Wechsler et al. 2006; Croton et al. 2007; Gao & White 2007; Zu et al. 2008; Dalal et al. 2008; Li et al. 2008; Lacerna & Padilla 2011, 2012; Zentner et al. 2014; Lacerna et al. 2014; Hearin et al. 2015). We will show that assembly bias is present in both EAGLE and SHAM galaxies, increasing the clustering amplitude by 20 % on scales from 2 to 11 Mpc. To our knowledge, this is the first detection of assembly bias in a hydrodynamical simulation. This result supports the idea that Halo Occupation Distribution (HOD) models (e.g., Seljak 2000; Peacock & Smith 2000; Scoccimarro et al. 2001), which are a phenomenological parametrization for the number of galaxies hosted by haloes of a given mass, introduce bias in the calculation of galaxy clustering when they assume that halo occupation is a function only of halo mass.

Finally, we will track the small residual differences in the clustering of SHAM and EAGLE galaxies to the failure of a key assumption of SHAM (as commonly implemented): for the same $V_{\rm relax}$, central and satellite subhaloes host the same galaxies independently of their host halo mass. We will find that this supposition is broken due to the influence of the environment and the star formation that satellite galaxies experience after having been accreted. Both effects correlate with the mass of the DM host, which suggests that future SHAM implementations that employ both host halo mass and $V_{\rm relax}$ could yield even more accurate predictions for the clustering signal.

Our paper is organized as follows. In §2.2 we describe the simulations, halo and galaxy catalogues, and merger trees that we use. In §2.3 we discuss different implementations of SHAM and introduce $V_{\rm relax}$, a new proxy for stellar mass. In §2.4 we analyse the accuracy with which SHAM can predict the galaxy satellite fraction, host halo mass, clustering, and assembly bias. We discuss the limitations of SHAM in §2.5. We conclude and summarize our most important results in §2.6.

2.2 Numerical Simulations

In this section we provide details of the main datasets that we employ. This includes a brief description of the numerical simulations, halo and galaxy catalogues, merger trees, and of a technique to identify the same structures in our hydrodynamical and gravity-only simulations.

2.2.1 The EAGLE suite

The simulations we analyse in this paper belong to the EAGLE project (Schaye et al. 2015; Crain et al. 2015) conducted by the Virgo consortium. EAGLE is a suite of high-resolution hydrodynamical simulations aimed at understanding the formation of galaxies in a cosmological volume. The runs employed a pressure-entropy variant (Hopkins 2013) of the Tree-PM smoothed particle hydrodynamics code GADGET3 (Springel 2005), the time step limiters of Durier & Dalla Vecchia (2012), and implement state-of-the-art subgrid physics (as described by Schaye et al. 2015), including metal-dependent radiative cooling and photo-heating (Wiersma et al. 2009a),

Table 2.1: EAGLE/DMO cosmological and numerical parameters. The cosmological parameter values are taken from Planck Collaboration et al. (2014a,b).

Parameter	EAGLE/DMO
$\Omega_{\rm m}$	0.307
Ω_Λ	0.693
$\Omega_{\rm b}$	0.04825
H_0[km s^{-1} Mpc^{-1}]	67.77
σ_8	0.8288
$n_{\rm s}$	0.9611
Max. proper softening [kpc]	0.70
Num. of baryonic particles	$1504^3/-$
Num. of DM particles	$1504^3/1504^3$
Initial baryonic particle mass [10^7M$_\odot$]	$0.181/-$
DM particle mass [10^7M$_\odot$]	$0.970/1.150$

Notes. $\Omega_{\rm m}$, Ω_Λ, and $\Omega_{\rm b}$ are the densities of matter, dark energy, and baryonic matter in units of the critical density at redshift zero. H_0 is the present day Hubble expansion rate, σ_8 is the linear fluctuation amplitude at 8 h^{-1} Mpc, and $n_{\rm s}$ is the scalar spectral index.

chemodynamics (Wiersma et al. 2009b), gas accretion onto supermassive black holes (Rosas-Guevara et al. 2015), star formation (Schaye & Dalla Vecchia 2008), stellar feedback (Dalla Vecchia & Schaye 2012), and AGN feedback.

The EAGLE suite includes runs with different physical prescriptions, resolutions, and volumes. Here, we study the largest simulation, which follows 1504^3 gas particles and the same number of DM particles inside a periodic box with a side length of 100 Mpc. The large volume and high resolution of this simulation are essential for a careful analysis of SHAM. The cosmological parameters used in EAGLE are those preferred by the analysis of *Planck* data (Table 2.1). This implies a gas particle mass equal to 1.81×10^6M$_\odot$ and a DM particle mass equal to 9.70×10^6M$_\odot$. We highlight that EAGLE is well suited to this study because it was calibrated to reproduce the galaxy stellar mass function at $z \sim 0$. The agreement with observations is especially good over the mass range that we will analyse here (fig. 4 of Schaye et al. 2015).

The 100 Mpc box was resimulated including only gravitational interactions and sampling the density field with 1504^3 particles of mass 1.15×10^7M$_\odot$. Hereafter, we refer to this simulation and its hydrodynamical counterpart as DMO and EAGLE, respectively. The cosmological and some of the numerical parameters employed in these simulations are provided in Table 2.1.

2.2.2 Catalogues and mergers trees

In each simulation, haloes were identified using only DM particles and a standard Friends-Of-Friends (FOF) group-finder with a linking parameter $b = 0.2$ (Davis et al.

1985). Gas and star particles are assigned to the same FOF halo as their closest DM particle. For each FOF halo we compute a spherical-overdensity mass, M_{200}, defined as the mass inside a sphere with mean density equal to 200 times the critical density of the Universe, $\rho_{\rm crit}(z)$;

$$M_{200} = \frac{4\pi}{3} 200 \rho_{\rm crit} r_{200}^3, \tag{2.1}$$

where r_{200} is the radius of the halo, $\rho_{\rm crit}(z) = \frac{3\,H^2(z)}{8\pi G}$, G is the gravitational constant, and $H(z)$ is the value of the Hubble parameter $H(z) = H_0\sqrt{\Omega_{\rm m}(1+z)^3 + \Omega_\Lambda}$.

Self-bound structures inside FOF haloes, termed subhaloes, were identified using all particle types and the SUBFIND algorithm (Springel et al. 2001; Dolag et al. 2009). Hereafter, we will refer to the subhalo located at the potential minimum of a given FOF halo as the "central", to any other structures as "satellites", and to subhaloes with more than one star particle as EAGLE "galaxies".

The position of each galaxy is assumed to be that of the particle situated at the minimum of the gravitational potential of the respective subhalo. The galaxy velocity is assumed to be that of the centre-of-mass of the subhalo [1]. The stellar mass, $M_{\rm star}$, is the total mass of all star particles linked to a given EAGLE galaxy. The gas mass ($M_{\rm gas}$) and the DM mass ($M_{\rm DM}$) are computed in the same manner but using gas or DM particles, respectively. We verified that our results are insensitive to the exact definition of $M_{\rm star}$: we repeated our analysis defining $M_{\rm star}$ as the mass inside a sphere of 20, 30, 40, 50, 70, or 100 kpc radius. We found that different mass definitions only produces sub-percent differences in the galaxy clustering.

We employ "merger trees" to follow the evolution of haloes and subhaloes, their mass growth, tidal stripping, mergers, as well as transient effects in their properties. Our trees were built using the algorithm described in Jiang et al. (2014), employing 201 snapshots for DMO and 29 snapshots for EAGLE. In both simulations the output times were approximately equally spaced in $\log(a)$ for $a > 0.2$, where a is the cosmic scale factor.

Finally, we note that to avoid problems related to subhalo fragmentation and spurious structures, we remove from our analysis satellites without resolved progenitors.

2.2.3 The EAGLE and DMO crossmatch

EAGLE and DMO share the same initial conditions, so we expect roughly the same non-linear objects to form in both simulations. This is a powerful feature: it enables us to identify the EAGLE galaxy that a given DMO subhalo is expected to host, and thus, to probe directly the assumptions of SHAM.

In practice, we link DMO subhaloes to EAGLE galaxies following the process described by Schaller et al. (2015); see also Velliscig et al. (2014). For every subhalo in EAGLE we select the 50 most-bound DM particles. If we find a subhalo in DMO which shares at least half of them, the link is made. We confirm the link if, repeating the same process starting from each DMO subhalo, we identify the same pair. We

[1] We checked that the mean difference between the bulk velocity of DM particles and star particles in the inner 30 kpc for the subhaloes with $8.77 < M_{\rm star}[{\rm M}_\odot] < 10.77$ is smaller than 10 km s^{-1}.

Table 2.2: Number of central and satellite EAGLE galaxies for four stellar mass bins. In parentheses we provide the percentage of EAGLE galaxies with a counterpart in DMO.

$\log_{10}(M_{\rm star}[M_\odot])$	EAGLE	
	Central	Satellites
$8.77 - 9.27$	3954 (92 %)	3475 (68 %)
$9.27 - 9.77$	2550 (92 %)	2068 (74 %)
$9.77 - 10.27$	1551 (94 %)	1247 (76 %)
$10.27 - 10.77$	968 (92 %)	652 (80 %)

only search the pairs with more than 174 DM particles in each simulation, which corresponds to a minimum halo mass of $2 \times 10^9 \mathrm{M}_\odot$ in DMO. This procedure yields a catalogue of 13687 galaxies with $10^{8.77} < M_{\rm star}[\mathrm{M}_\odot] < 10^{10.77}$.

In Table 2.2 we list the fraction of successfully matched centrals and satellites, for four stellar mass bins. Overall, the match is successful for more than 90 % of centrals in EAGLE, independently of their mass. The success rate drops to $68 - 80$ % for satellites, with low-mass satellites showing the lowest percentage. This is a consequence of the finite mass resolution of the simulations (see also Appendix A), the mass loss due to interactions with the host halo, small differences in the timing at which mergers happen, and the high-density environment in which they reside.

2.3 Subhalo abundance matching

In this section we discuss different SHAM flavours and their implementation in DMO.

2.3.1 SHAM flavours

The main assumption of SHAM is that there is a one-to-one relation between a property of a DM subhalo and a property of the galaxy that it hosts. The galaxy property is usually taken to be the stellar mass (or K-band luminosity), since this is expected to be tightly correlated with the DM content of the host halo (contrary to e.g. the star formation rate, which could be more stochastic). The subhalo property should capture the time-integrated mass of gas available to fuel star formation, but there is no consensus as to what the most adequate subhalo property is[2].

A commonly used property in SHAM is the maximum of the radial circular velocity profile (which can be regarded as a measure of the depth of the potential well of a

[2]Properties used in the literature include $M_{\rm DM}$ (Vale & Ostriker 2004; Shankar et al. 2006), maximum circular velocity at present for centrals and at infall for satellites (Conroy et al. 2006), virial mass for centrals and mass at infall for satellites (Wetzel & White 2010; Behroozi et al. 2010), virial mass for centrals and the highest mass along the merger history for satellites (Moster et al. 2010), and highest circular velocity along the merger history (Trujillo-Gomez et al. 2011; Nuza et al. 2013) (see Reddick et al. 2013, for a detailed comparison between the previous properties).

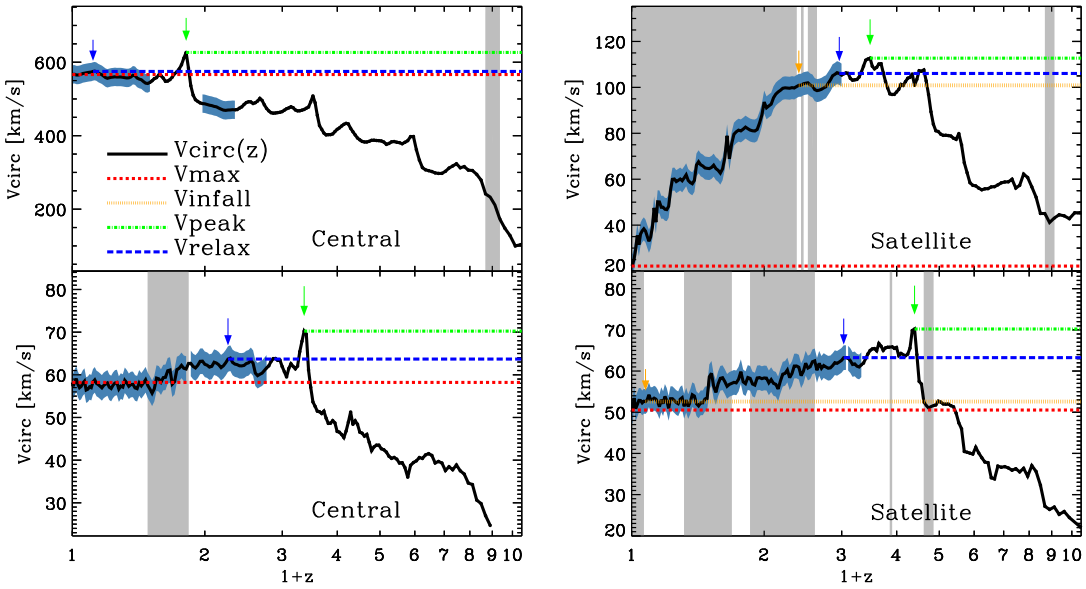


Figure 2.1: Evolution of the maximum circular velocity of two central (left panel) and two satellite (right panel) subhaloes in DMO. The black solid lines show the circular velocity, the grey coloured areas the periods during which the subhaloes are satellites, and the blue coloured regions the intervals during which the subhaloes satisfy our relaxation criterion. Horizontal lines highlight the circular velocity at $z = 0$ ($V_{\rm max}$, red dashed line), the circular velocity at the last infall for satellites and $V_{\rm max}$ for centrals ($V_{\rm infall}$, orange dotted line), the maximum circular velocity that a subhalo has had ($V_{\rm peak}$, green dot-dashed line), and the maximum circular velocity that a subhalo has reached while it satisfied our relaxation criterion ($V_{\rm relax}$, blue long dashed line).

subhalo) defined at a suitable time:

$$V_{\rm circ}(z) \equiv \max\left[\sqrt{GM(z,<r)/r}\right]. \tag{2.2}$$

where $M(<r)$ is the mass enclosed inside a radius r.

There are several reasons to prefer circular velocity over halo mass in SHAM: i) it is typically reached at one tenth of the halo radius, so it is a better characterisation of the scales that we expect to affect the galaxy most directly; ii) it is less sensitive to the mass stripping that a halo/subhalo experiences after it has been accreted by a larger object (Hayashi et al. 2003; Kravtsov et al. 2004; Nagai & Kravtsov 2005; Peñarrubia et al. 2008); iii) it does not depend on the definition of halo/subhalo mass.

However, the $V_{\rm circ}(z)$ of DM objects are complicated functions, which can display non-monotonic behaviour in time, with transient peaks and dips, and that are subject to environmental and numerical effects. This is illustrated by Fig. 2.1, which shows examples of the evolution of the circular velocity for two central (left panel) and two satellite (right panel) subhaloes in DMO. These subhaloes are selected to illustrate

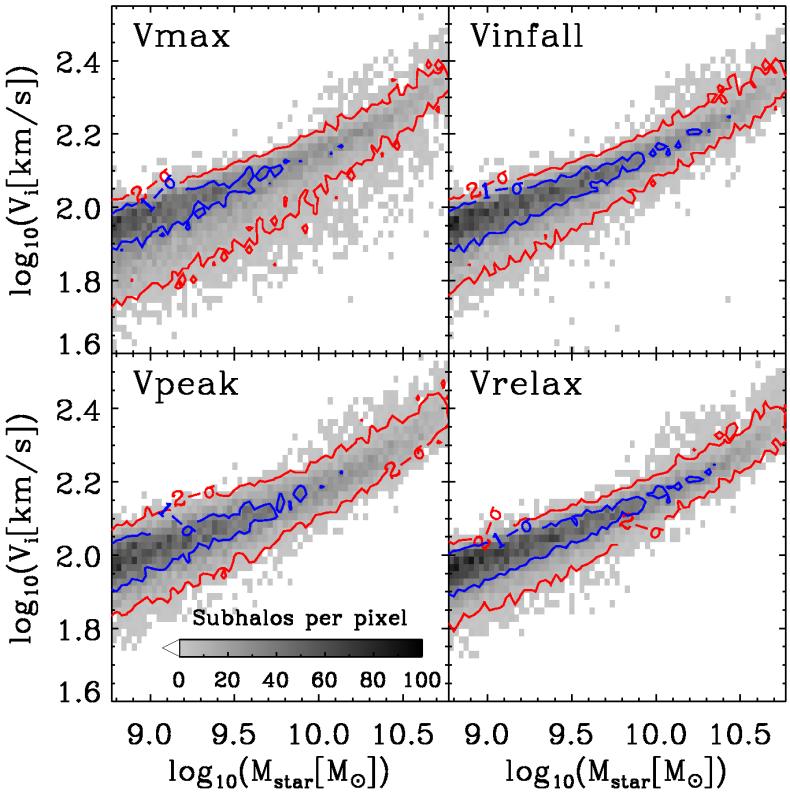


Figure 2.2: Relation between $M_{\rm star}$ of EAGLE galaxies and SHAM flavours for the corresponding DMO subhaloes. The grey scale represents the number of subhaloes per pixel, which ranges from 1 (light grey) to 100 (black). Blue and red contours mark the regions containing 68% and 95% of the distribution, respectively.

the evolution of the maximum circular velocity in typical centrals and satellites. We can see that there is no obvious time at which $V_{\rm circ}(z)$ should be computed for an accurate SHAM.

We will implement four "flavours" of SHAM, each using $V_{\rm circ}(z)$ defined at a different time: $V_{\rm max}$, $V_{\rm peak}$, $V_{\rm infall}$, and $V_{\rm relax}$ (each marked by horizontal lines and arrows of a different colour in Fig. 2.1). The first three flavours have been used previously in the literature, whereas the fourth is first used in this work. We discuss the four SHAM flavours next.

1) $V_{\rm max}$ is the maximum circular velocity of a subhalo at the present time, $V_{\rm circ}(z = 0)$.

2) $V_{\rm infall}$ is the maximum circular velocity at the last time a subhalo was identified as a central.

3) $V_{\rm peak}$ is the maximum circular velocity that a subhalo has reached.

4) $V_{\rm relax}$ is the maximum circular velocity that a subhalo has reached during the periods in which it satisfied a relaxation criterion. The criterion we use is

$\Delta t_{\rm form} > t_{\rm cross}$, following a similar approach to Ludlow et al. (2012). The motivation is that after a major merger, DM haloes typically need of the order of one crossing time ($t_{\rm cross} = 2\,r_{200}/V_{200} = 0.2/H(z)$) to return to equilibrium. Thus, we define $\Delta t_{\rm form}$ as the look-back time from a given redshift $z_{\rm i}$ to the redshift where the main progenitor of a subhalo reached 3/4 of the subhalo mass at $z_{\rm i}$ (we tested other definitions for the formation time, from 4/5 to 1/2, finding roughly the same results). The periods during which this condition is satisfied are shown as blue shaded regions in Fig. 2.1. We can compute $V_{\rm relax}$ for more than the 99% of the subhaloes in DMO and we remove the subhaloes where $V_{\rm relax}$ cannot be calculated. We cannot compute $V_{\rm relax}$ for the full sample because this quantity is not defined for subhaloes younger than one crossing time.

Although $V_{\rm circ}$ should generally not be affected by the stripping of the outer layers of a halo, in the right panel of Fig. 2.1 we can see that it does still evolve for satellites. The decrease in $V_{\rm circ}(z)$ after infall is in large part due to tidal heating, a process which reduces the density in the inner regions of the satellites (Gnedin 2003; Hayashi et al. 2003; Kravtsov et al. 2004). The tidal heating is related to the position of a subhalo inside its host halo, being maximum at pericentric passages. We can see an extreme case of tidal interactions in the top right panel, where this subhalo has lost more than 99% of its mass since it became a satellite. After the last infall at $1+z \sim 2.3$ (grey shaded region), the value of $V_{\rm circ}$ decreased by about 80% in a series of steps ($z \sim 1$, 0.5, 0.3, 0.1, 0.05, and 0), which indeed coincide with pericentric passages. This implies that satellite galaxies have lower values of $V_{\rm max}$ than central galaxies of the same stellar mass. Thus, a $V_{\rm max}$-based SHAM will underestimate the fraction of satellites.

Tidal heating and stripping affect not only satellites but also "backsplash satellites", i.e. centrals at $z = 0$ which were satellites in the past, reducing their circular velocity while they were inside a larger halo. An example of this process is shown in the bottom left panel of Fig. 2.1, where the circular velocity of this subhalo was reduced by about 7% in the period during which it was a satellite (while the mass was reduced by 50%).

$V_{\rm infall}$ is less affected by these problems. Unfortunately, this parameter also underestimates $V_{\rm circ}$ for satellites because tidal heating starts to act even before a satellite is accreted by its future host halo (Kravtsov et al. 2004; Wetzel et al. 2013, 2014). This can be seen in the top (bottom) right panel Fig. 2.1, where the value of $V_{\rm circ}$ starts to decrease at $1+z \sim 3.4$ ($1+z \sim 4.4$) while the subhalo is accreted at $1+z \sim 2.4$ ($1+z \sim 1.2$).

Additionally, there are new problems associated with $V_{\rm infall}$. The first concerns satellite-satellite mergers (Angulo et al. 2009; Wetzel et al. 2009), which should increase the mass of stars in a satellite but this is not captured by $V_{\rm infall}$. The second is related to the definition of $V_{\rm infall}$; it is not clear whether we should consider $V_{\rm infall}$ as the circular velocity at the last infall or at previous accretion events. We can see in the bottom right panel of Fig. 2.1 a satellite which has undergone several alternating central/satellite periods, decreasing in total its circular velocity by 20% and its mass by 70%.

An alternative solution is provided by $V_{\rm peak}$ since it can capture all episodes during which the subhalo grows, and it is not affected by a reduction of $V_{\rm circ}$ due to envi-

Table 2.3: Parameters of the functions that fit the mean, μ, and standard deviation, σ, of the model for $P(\log_{10} M_{\rm star}[{\rm M}_\odot]|\log_{10} V_i[{\rm km\ s^{-1}}])$. The unit of V_i is km s^{-1}.

	$\sigma = a + b \log_{10} V_i$		$\mu = a + b \tan^{-1}(c + d \log_{10} V_i)$			
	a	b	a	b	c	d
$V_{\rm max}$	0.60	-0.20	7.03	5.52	-1.84	1.12
$V_{\rm infall}$	0.53	-0.16	7.01	5.52	-1.84	1.12
$V_{\rm peak}$	0.55	-0.16	7.70	5.42	-1.89	1.05
$V_{\rm relax}$	0.59	-0.20	7.14	5.55	-1.86	1.10

ronmental effects. However, this definition similarly has its own problems. During periods of rapid mass accretion, DM haloes are usually out of equilibrium (Neto et al. 2007). In particular, during major mergers the concentration can be artificially high (this is a maximum compression phase of halo formation), which temporarily increases the value of $V_{\rm circ}$ (e.g. Ludlow et al. 2012; Behroozi et al. 2014). This effect is responsible for the peaks seen in all four panels of Fig. 2.1. Although at any given time it is rare to find a halo in this phase, the value of $V_{\rm peak}$ will likely be assigned during one of these phases, and will thus overestimate the depth of the potential well. In addition, this effect makes the predictions of $V_{\rm peak}$ dependent on the number and intervals of the output times of a given simulation.

Here we propose a new measure, $V_{\rm relax}$, designed to overcome the problems of $V_{\rm max}$, $V_{\rm infall}$, and $V_{\rm peak}$. It is marked by arrows and horizontal lines of blue colour in Fig. 2.1. $V_{\rm relax}$ is insensitive to tidal heating, transient peaks, and consistently defined for centrals, satellites, and backsplash satellites. We emphasise that it is desirable to eliminate the aforementioned problems because they represent changes in $V_{\rm circ}$ which are not expected to correlate with the growth history of $M_{\rm star}$, and will thus add extra noise to SHAM.

We now take a first look at the performance of each SHAM flavour. Fig. 2.2 shows the relation between each of the four properties described above for DMO subhaloes, as indicated by the legend, and $M_{\rm star}$ of their galaxy counterpart in EAGLE (see §2.2.3). All panels show a tight correlation, which supports the main assumption of SHAM, that the relation between stellar mass and SHAM parameters should be monotonic. However, the scatter in the relation is different in each panel because of the effects discussed in this section: $V_{\rm max}$ shows the largest and $V_{\rm relax}$ the smallest dispersion. In the next sections we will quantify the performance of each SHAM flavour in detail.

2.3.2 Implementation

The first step to implement the four flavours of SHAM is to compute $P(\log_{10} {\rm M_{star}}|\log_{10} V_i)$: the probability that a subhalo hosts a galaxy of mass $M_{\rm star}$ given a certain value of the SHAM flavour V_i. We compute this quantity as follows:

1) We select subhalo-galaxy pairs from the matched catalogues (see §2.2.3) with $\log_{10} M_{\rm star}[{\rm M}_\odot] > 7$ and divide them according to $\log_{10} V_i$ in bins of 0.05 dex.

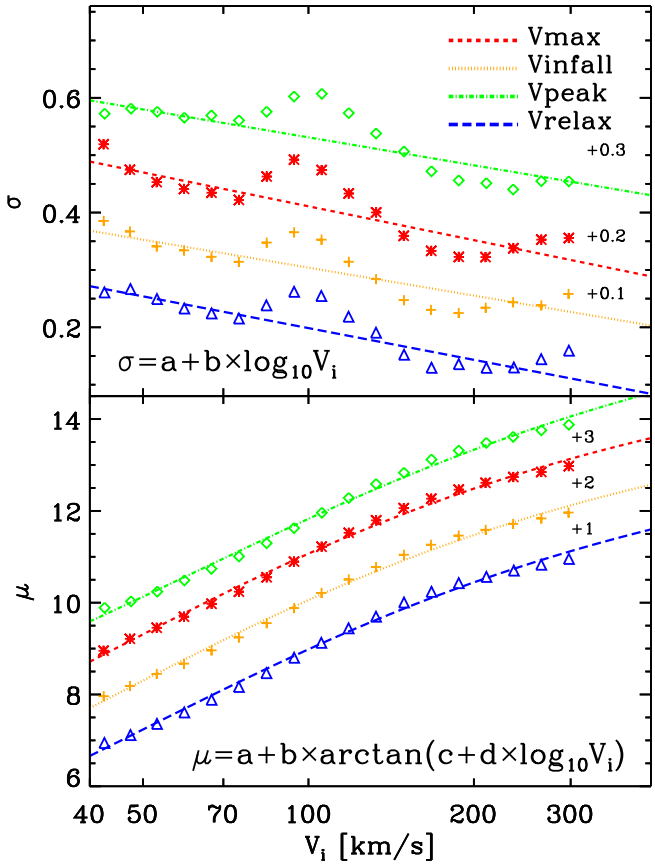


Figure 2.3: Standard deviation (top panel) and mean (bottom panel) of the Gaussians used to fit Probability Distribution Functions (PDF)text for $\log_{10} M_{\rm star}[{\rm M}_\odot]$. For clarity, we have shifted the σ (μ) of $V_{\rm max}$, $V_{\rm infall}$, and $V_{\rm peak}$ by +0.3, +0.2, and +0.1 (+3, +2, and +1), respectively. The best fitting functions are shown by coloured lines, and the values of the respective parameters are given in Table 2.3.

We discard bins with fewer than 100 objects.

2) For each $\log_{10} V_i$ bin, we compute the distribution of $\log_{10} M_{\rm star}$ and fit it by a Gaussian function, $G \sim \exp(-0.5(\log_{10} M_{\rm star} - \mu)^2/(\sigma)^2)$, where μ is the mean and σ the dispersion.

3) We fit a linear function, $\sigma = a + b \log_{10} V_i$, to $\sigma(\log_{10} V_i)$ and an arctangent, $\mu = a + b \tan^{-1}(c + d \log_{10} V_i)$, to $\mu(\log_{10} V_i)$. The values of the best-fit parameters are given in Table 2.3 and the quality of the fit can be judged from Fig. 2.3.

4) Using these functions, we model $P(\log_{10} {\rm M}_{\rm star} | \log_{10} V_i)$ as $G[\mu(\log_{10} V_i), \sigma(\log_{10} V_i)]$.

Our second step is to assign a value of $M_{\rm star}$ to every subhalo in DMO (not only those with an EAGLE counterpart) by randomly sampling $P(\log_{10} {\rm M}_{\rm star} | \log_{10} V_i)$. This creates a catalogue that captures the appropriate stochastic relation between $M_{\rm star}$ and the parameter V_i. If the relation for EAGLE galaxies were also stochastic

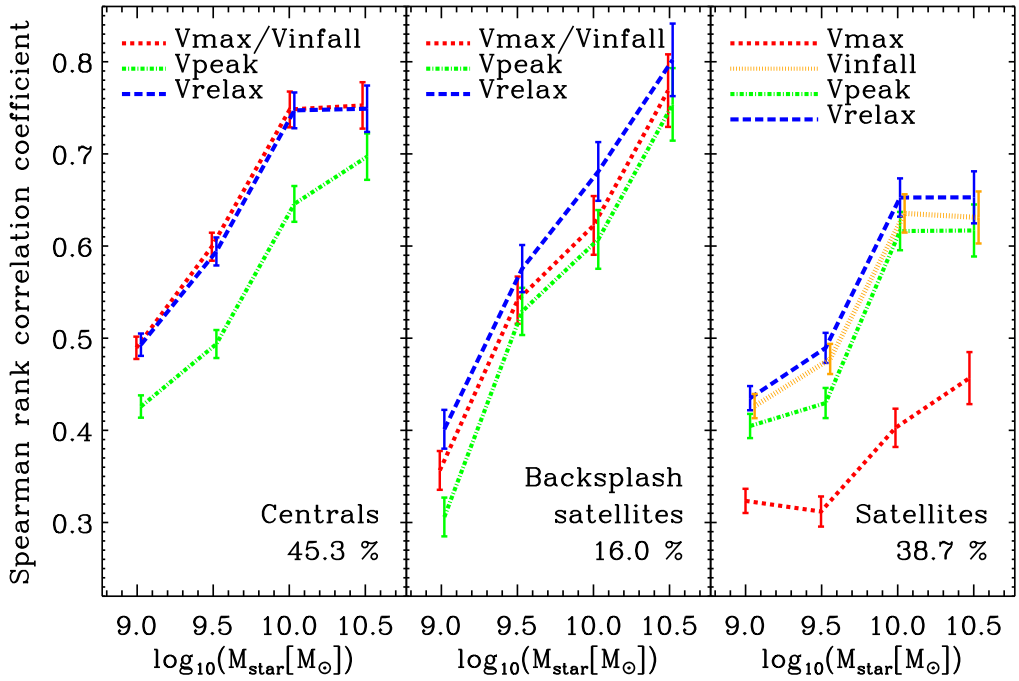


Figure 2.4: The Spearman rank correlation coefficient between the $M_{\rm star}$ of EAGLE galaxies and each of four parameters used to perform SHAM. The subhaloes are divided into three categories: centrals (left panel), backsplash satellites (central panel), and satellites (right panel), see the main text for more details. The fraction of objects in each category is given in the legend. The red (orange) points are displaced horizontally by -0.03 (+0.03) dex for clarity.

with respect to the underlying density field, then we would expect these catalogues to have the same clustering properties as EAGLE.

We note we have verified that the resulting stellar mass function agrees closely with that of the EAGLE simulation. However, to ensure *identical* mass functions and thus to make subsequent comparisons more direct, we assign to each SHAM galaxy the value of $M_{\rm star}$ of the EAGLE galaxy at the same rank order position. Hereafter, we will refer generically to the galaxy catalogues created in this way as "SHAM galaxies" and specifically to the galaxy catalogues generated by a particular SHAM parameter as "V_i galaxies".

We compute 100 realizations of SHAM for every flavour using different random seeds. The results presented in the following sections are the mean of all the realizations and the errors the standard deviation.

2.4 Results

In this section we test how well SHAM reproduces different properties of EAGLE galaxies. In particular, we will explore the predicted stellar mass of individual subhaloes

(§2.4.1), the HOD (§2.4.2), the number density profiles inside haloes (§2.4.2), the clustering in real and redshift space (§2.4.3, §2.4.3), and the assembly bias (§2.4.3).

We present results for 4 bins in stellar mass, as indicated in Table 2.2. This range was chosen to include only well sampled and well resolved galaxies (comprised of more than 230 star particles) and bins with enough galaxies to allow statistically significant analyses (more than 1600 galaxies).

2.4.1 Correlation between $M_{\rm star}$ and V_i

In §2.3 we discussed that in some cases $V_{\rm max}$, $V_{\rm infall}$, and $V_{\rm peak}$ are unintentionally affected by physical and numerical effects, which degrades the performance of SHAM. We also argued that $V_{\rm relax}$ does not present any obvious problem and thus we expected it to be the SHAM flavour that correlates most strongly with $M_{\rm star}$. This was qualitatively supported by Fig. 2.2. We start this section by quantifying these statements using the Spearman rank correlation coefficient between the $M_{\rm star}$ of EAGLE galaxies and the SHAM flavours of DMO subhaloes.

The Spearman coefficient measures the statistical dependence between two quantities and is defined as the Pearson correlation coefficient between the ranks of sorted variables. A value of unity implies a perfect correlation, which in our case means that the stellar mass of a galaxy is completely determined by its SHAM parameter, i.e. that the relation is monotonic and thus without scatter. A value close to zero means that the relation between the SHAM parameter and $M_{\rm star}$ is essentially random.

In Fig. 2.4 we show the Spearman coefficient for the correlation between $M_{\rm star}$ and each of our four SHAM parameters. We divide our sample into three groups: i) present-day central subhaloes that have been centrals for their entire merger history except for at most 4 snapshots (centrals, left panel), ii) present-day central subhaloes that have been satellites more than 4 snapshots in the past (backsplash satellites, central panel), and iii) present-day satellites (satellites, right panel).

In general, we find that the correlation increases with $M_{\rm star}$, that it is stronger for centrals than for satellites, and that $V_{\rm relax}$ displays the strongest correlation with $M_{\rm star}$. Regarding the different SHAM flavours, we find that i) for centrals $V_{\rm peak}$ produces the weakest correlation, ii) for satellites $V_{\rm max}$ shows the weakest correlations, and iii) $V_{\rm infall}$ and $V_{\rm relax}$ consistently display the best performance, with $V_{\rm relax}$ showing a slight improvement over $V_{\rm infall}$ for satellites.

Our results can be understood from the discussion in §2.2. For centrals, $V_{\rm max}$ and $V_{\rm infall}$ are identical by construction and they are close to the value of $V_{\rm relax}$ because $V_{\rm circ}$ tends to increase with decreasing redshift for centrals. On the other hand, $V_{\rm peak}$ is usually established while $V_{\rm circ}$ is temporarily enhanced as a result of merger events. For backsplash satellites, $V_{\rm max}$ and $V_{\rm infall}$ are also identical by construction, but, unlike $V_{\rm relax}$, they are insensitive to their more complicated history, which explains their weaker correlation with $M_{\rm star}$.

Finally, satellites display the weakest correlations, with $V_{\rm max}$ presenting the lowest correlation coefficient. This is because $V_{\rm circ}$ decreases soon after infall, whereas the stellar mass can still grow until the gas is completely exhausted (although tidal forces may strip stars). $V_{\rm infall}$ alleviates this problem but the interaction between the satellites and their host haloes starts before the satellites reach the virial radii of their host

Table 2.4: Satellite fraction for EAGLE and SHAM galaxies using $V_{\rm max}$, $V_{\rm infall}$, $V_{\rm peak}$, and $V_{\rm relax}$.

$\log_{10}(M_{\rm star}[{\rm M}_\odot])$	$V_{\rm max}$	$V_{\rm infall}$	$V_{\rm peak}$	$V_{\rm relax}$	EAGLE
	Satellite fraction				
$8.77 - 9.27$	0.32	0.43	0.46	0.45	0.47
$9.27 - 9.77$	0.30	0.42	0.44	0.43	0.45
$9.77 - 10.27$	0.28	0.40	0.41	0.41	0.44
$10.27 - 10.77$	0.25	0.37	0.38	0.37	0.40

haloes (Hayashi et al. 2003; Bahé et al. 2013). Because of this, $V_{\rm relax}$ better captures the expected evolution in $M_{\rm star}$. Lastly, $V_{\rm peak}$ is still affected by the out-of-equilibrium artefacts discussed above.

In sections §2.4.2 and §2.4.3 we will investigate how the different correlations impact the predictions for the clustering of EAGLE galaxies.

2.4.2 The properties of SHAM galaxies

To predict the correct galaxy clustering, SHAM has to associate galaxies with the correct subhaloes, to allocate the right proportion of centrals and satellites, and to place galaxies following the correct radial distribution. Therefore, before presenting our results regarding the clustering, we will explore these ingredients separately.

Halo occupation distribution

The panels of Fig. 2.5 show the distribution of host halo masses for centrals and satellites in different $M_{\rm star}$ bins. The left (right) curves display the number of centrals (satellites) in haloes of a given mass multiplied by the linear bias[3] expected for haloes of that mass and normalized by the total number of subhaloes. The quantity plotted can be interpreted as the relative contribution to the large-scale clustering from galaxies hosted by haloes of different mass. In each panel, the histogram presents the results for EAGLE galaxies and the coloured lines the results of the SHAM implementations detailed in §2.3.2. For EAGLE galaxies we employ the M_{200} of their host halo DMO counterpart, which makes this plot less sensitive to baryonic effects that might systematically change the mass of DM haloes. For the 5.1% of EAGLE galaxies hosted by a halo without DMO counterpart, we multiply M_{200} by $f_{\rm DM} = 1 - (\Omega_{\rm b}/\Omega_{\rm m}) = 0.843$. This is the average difference in M_{200} between the hydrodynamic and gravity-only EAGLE simulations, as reported by Schaller et al. (2015).

Firstly, we see that using $V_{\rm max}$ as SHAM parameter results in shifted M_{200} distributions and an underprediction, of about 30 %, of the number of satellites for all $M_{\rm star}$

[3]We calculate the linear bias as $b = 1 + \frac{\nu^2 - 1}{\delta_{\rm c}}$ (Mo & White 1996), where $\delta_{\rm c} \approx 1.69$ is the critical linear overdensity at collapse and $\nu = \delta_{\rm c}/\sigma(M, z)$ is the dimensionless amplitude of fluctuations which produces haloes of mass M at redshift z.

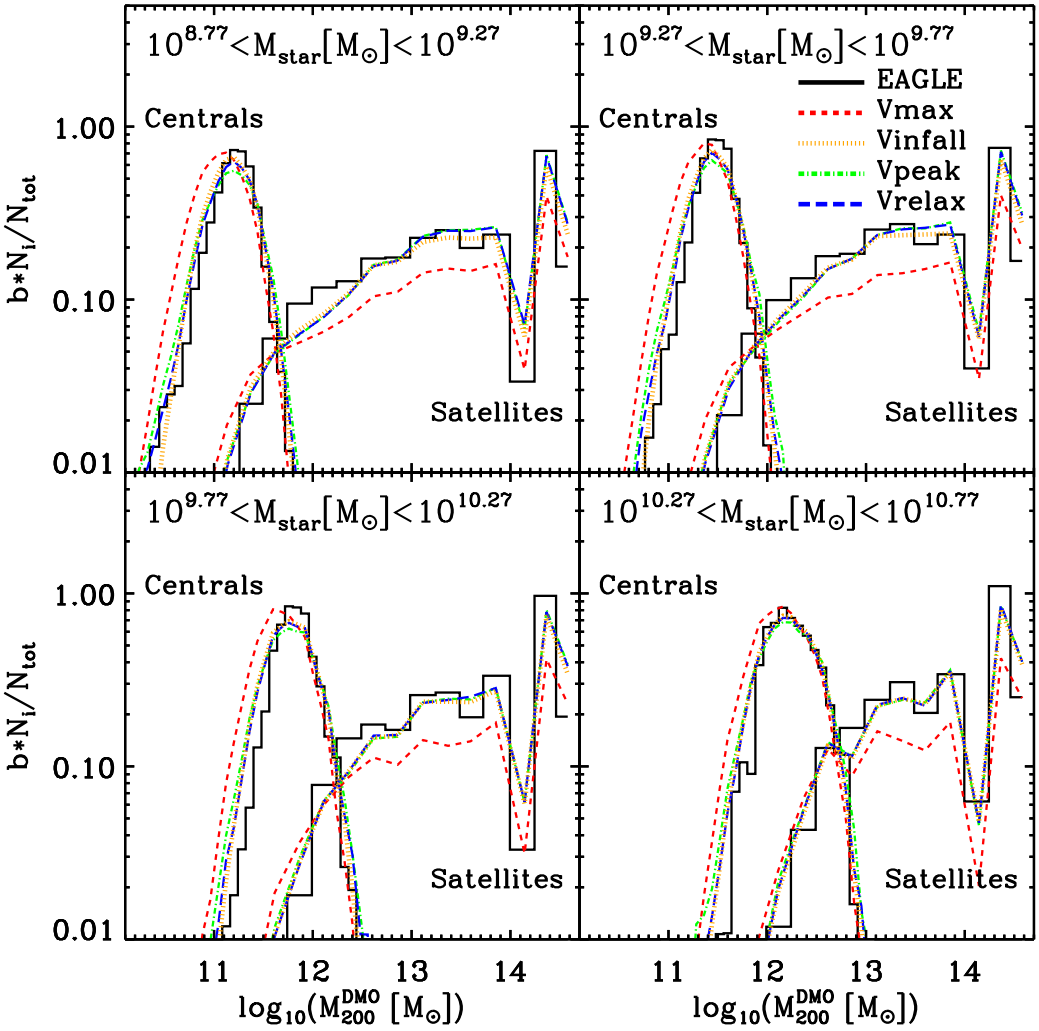


Figure 2.5: The distribution of host halo masses, M_{200}, for SHAM and EAGLE galaxies in different $M_{\rm star}$ bins. Histograms show the results for EAGLE galaxies and coloured lines for different SHAM flavours, as detailed in §2.3.2. The left (right) curves display the number N_i of centrals (satellites) in haloes of a given mass multiplied by the linear bias b and normalized by the total number of subhaloes N_{tot}. Therefore, the y-axis reflects the relative contribution of galaxies in different host halo mass bins to the large-scale correlation function. Note that for EAGLE galaxies we employ the M_{200} of the DMO counterpart, which makes our comparison less dependent on the baryonic processes which might alter the mass of the host halo.

Table 2.5: Number of satellites as a function of $M_{\rm star}$ and M_{200} for EAGLE and SHAM galaxies using $V_{\rm relax}$.

$\log_{10}(M_{\rm star}[{\rm M}_\odot])$	$\log_{10}(M_{200}[{\rm M}_\odot])$	EAGLE	$V_{\rm relax}$
		N. of satellites	
$8.77 - 9.27$	$11.6 - 12.6$	1060	780
	$12.6 - 13.6$	1274	1328
	$13.6 - 14.6$	945	1057
$9.27 - 9.77$	$11.6 - 12.6$	584	444
	$12.6 - 13.6$	834	838
	$13.6 - 14.6$	633	695
$9.77 - 10.27$	$11.6 - 12.6$	293	208
	$12.6 - 13.6$	495	482
	$13.6 - 14.6$	459	452
$10.27 - 10.77$	$11.6 - 12.6$	65	61
	$12.6 - 13.6$	280	253
	$13.6 - 14.6$	307	292

bins. This is a consequence of the reduction of $V_{\rm max}$ for satellites after being accreted, which introduces centrals hosted by lower-mass haloes into the SHAM sample.

The distribution of EAGLE galaxies is closely reproduced by the other SHAM implementations, for all stellar mass bins. The distributions for central galaxies have almost identical shapes and peak at roughly the same host halo mass. Note, however, that compared to $V_{\rm infall}$ and $V_{\rm relax}$, $V_{\rm peak}$ yields systematically broader distributions for centrals. This is consistent with the differences in the correlation coefficient shown in the left panel of Fig. 2.4.

Additionally, the $V_{\rm infall}$, $V_{\rm peak}$, and $V_{\rm relax}$ satellite fractions agree to within $\sim 5\%$ with those in EAGLE, although they are systematically lower, as shown in Table 2.4. However, for the two lowest stellar mass bins, there is a slight overestimate of the number of satellites in haloes of mass $M_{200} > 10^{13}{\rm M}_\odot$, and a somewhat larger underestimate for haloes of mass $M_{200} < 10^{13}{\rm M}_\odot$, as Table 2.5 shows. Since the difference is greater for the high-mass haloes, the overall satellite fraction is underestimated. We will analyse the repercussion of these small differences in forthcoming sections.

Radial distribution of satellites

Fig. 2.6 shows the spherically averaged number density profiles of satellite galaxies with $8.77 < \log_{10} M_{\rm star}[{\rm M}_\odot] < 10.77$, normalized to the mean number density within r_{200}. We show results for galaxies inside haloes in three DMO halo mass bins, as indicated by the legend. The data points represent the profiles measured using EAGLE galaxies, whereas coloured lines display the stacked results for SHAM galaxies. For comparison, we also plot the best-fit NFW profile to the EAGLE data, which appears

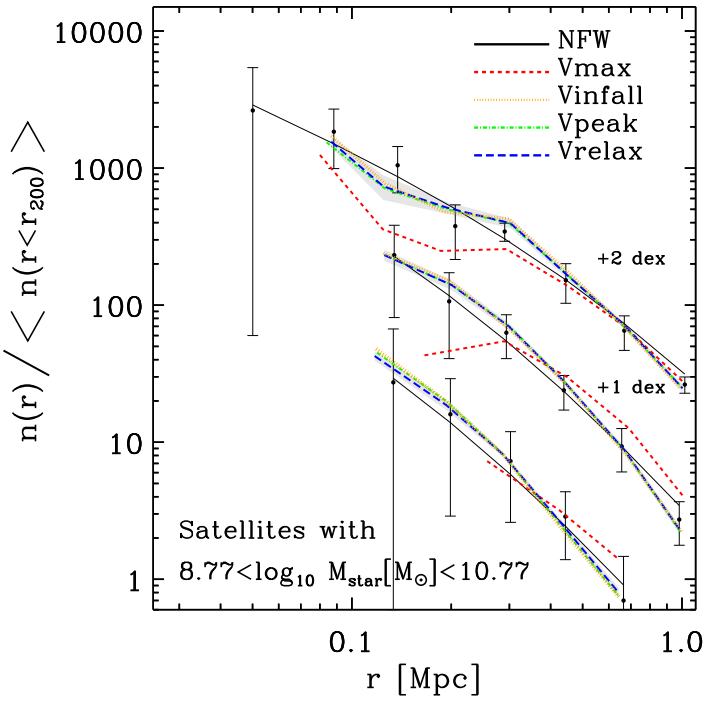


Figure 2.6: The radial distribution of galaxies with $8.77 < \log_{10}(M_{\rm star}[{\rm M}_\odot]) < 10.77$, inside haloes of mass $10^{13.0} - 10^{13.5}{\rm M}_\odot$, $10^{13.5} - 10^{14.0}{\rm M}_\odot$ (displaced by +1 dex), and more massive than $10^{14.0}{\rm M}_\odot$ (displaced by +2 dex). We present the spherically averaged number density, normalized to the mean number density within the host halo. Black symbols show the results for EAGLE galaxies, whereas coloured lines show stacked results from 100 realizations of SHAM using $V_{\rm max}$, $V_{\rm infall}$, $V_{\rm peak}$, and $V_{\rm relax}$. The error bars indicate the 1 σ scatter for EAGLE galaxies. The shaded region marks the standard deviation of 100 realizations of SHAM using $V_{\rm relax}$. We overplot the Navarro-Frenk-White (NFW) profiles (with $r_{\rm s} = 0.81$, 0.29, 0.21 Mpc from the most to the least massive halo sample) that best fit the EAGLE data points shown.

to be a good description over the range of scales probed.

Given the statistical uncertainties, the number density profiles of EAGLE and SHAM galaxies agree reasonably well with the exception of $V_{\rm max}$. For $V_{\rm max}$, the differences are greater, it predicts shallower profiles and a lack of objects in the inner parts compared to EAGLE. This is consistent with the effects described previously: the inner parts of haloes experience large tides and are also populated by the oldest subhaloes. In contrast, on scales $r > 0.1$ Mpc, the $V_{\rm peak}$, $V_{\rm infall}$ and $V_{\rm relax}$ profiles are consistent with the measurements from EAGLE for all three halo mass bins.

2.4.3 Galaxy clustering

We are now in the position to investigate the performance of SHAM in predicting the clustering of galaxies. We first discuss the Two Point Correlation Function (2PCF) in real-space (§2.4.3), then the monopole of the redshift-space correlation function

(§2.4.3), and we end with an exploration of assembly bias in both EAGLE and SHAM (§2.4.3).

We compute the 2PCF, $\xi(r)$, by Fourier transforming the galaxy number density field, which is a faster alternative to a direct pair count. We provide details of the procedure in Appendix B. We estimate the statistical uncertainties in the 2PCF of EAGLE galaxies using a spatial jack-knife resampling (e.g., Zehavi et al. 2005). Summarizing, we divide the simulation box in 64 smaller boxes and then we compute 64 2PCFs removing one of the small boxes each time. The statistical errors are the standard deviation of the 64 2PCFs. On the other hand, we assign errors to the 2PCF of SHAM galaxies by computing the standard deviation of 100 realizations for each SHAM flavour.

Real-Space Correlation Function

In Fig. 2.7 we compare the 2PCF for EAGLE galaxies (black solid line) with results of stacking 100 realizations of SHAM for different stellar mass bins. In the bottom panel of each subplot, we display the relative difference of the 2PCFs of each V_i galaxy sample and EAGLE ($\Delta\xi_i = \xi_i/\xi_{\rm EAGLE} - 1$).

Fig. 2.7 shows that $V_{\rm max}$ clearly underestimates the clustering on small scales, which is consistent with the underestimation of the satellite fraction discussed earlier. A lower satellite fraction also implies a lower mean host halo mass and a smaller bias, which explains the underestimation of the correlation function on larger scales.

On the other hand, $V_{\rm infall}$, $V_{\rm peak}$, and $V_{\rm relax}$ galaxies agree very closely with the EAGLE measurements. On scales greater than 2 Mpc, all three flavours are statistically compatible with the full hydrodynamical results. We note that the small differences are of the same order as the variance introduced by different samplings of $P(\log_{10}{\rm M_{star}}|\log_{10}V_i)$. For the two higher stellar mass bins, the statistical agreement is extended down to 400 kpc.

For the two lower stellar mass bins, we measure statistically significant differences on small scales, especially for $V_{\rm peak}$ and $V_{\rm relax}$ galaxies. The SHAM clustering appears to be $20-30$ % high, which could originate from either more concentrated SHAM galaxy distributions inside haloes, or from an excess of satellite galaxies. At first sight, the latter explanation appears to contradict our previous finding that the satellite fraction is underpredicted by SHAM. However, the small-scale clustering will be dominated by satellites inside very massive haloes[4], whose number is indeed overpredicted (c.f. Table 2.5).

Additionally, Fig. 2.5 showed that $V_{\rm infall}$ resulted in the same underestimation of the overall satellite fraction as $V_{\rm peak}$ and $V_{\rm relax}$ but a somewhat smaller satellite fraction in the high halo mass range. This explains the weaker small-scale clustering seen in Fig. 2.7 and consequently the slightly better agreement with EAGLE. Note, however, that the smaller number of satellites could be caused by the fact that $V_{\rm circ}$ decreases even before accretion, especially near very massive haloes. This suggests that the apparent improved performance of $V_{\rm infall}$ could be simply a coincidence. We

[4]For instance, in the case of the small-scale clustering of galaxies in the lowest stellar mass bin, the contribution of satellites inside haloes with $M_{200} > 10^{13}M_\odot$ is almost an order of magnitude larger than that of satellites in haloes with $M_{200} < 10^{13}M_\odot$.

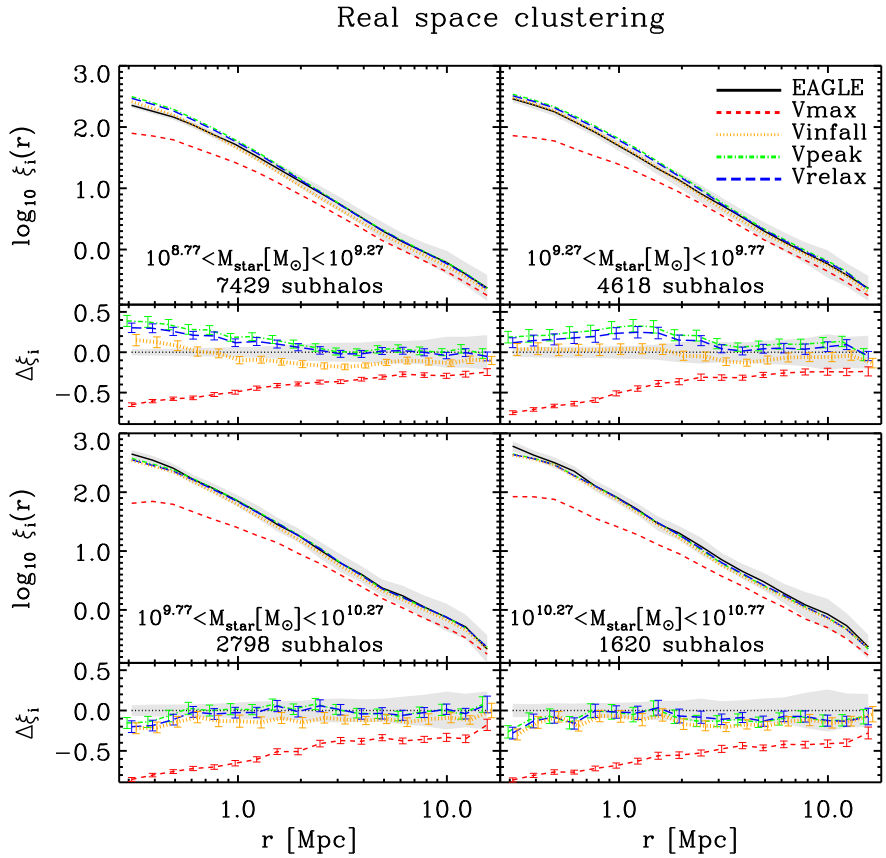


Figure 2.7: Real-space two-point correlation function for galaxies in different stellar mass bins. The black solid line shows the clustering in EAGLE, with the grey shaded region the jackknife statistical error. The coloured lines show the clustering predictions of SHAM using $V_{\rm max}$ (red dashed), $V_{\rm infall}$ (orange dotted), $V_{\rm peak}$ (green dot-dashed), and $V_{\rm relax}$ (blue long dashed). The error bars indicate the standard deviation of 100 realizations of SHAM for each flavour. In the lower half of each panel we display the relative difference of SHAM with respect to EAGLE ($\Delta\xi_i = \xi_i/\xi_{\rm EAGLE} - 1$). Note that the green and orange lines are slightly displaced horizontally for clarity. Using $V_{\rm relax}$ as SHAM parameter, we retrieve the clustering of EAGLE galaxies to within 10% on scales greater than 2 Mpc.

will investigate these hypotheses in §2.5.

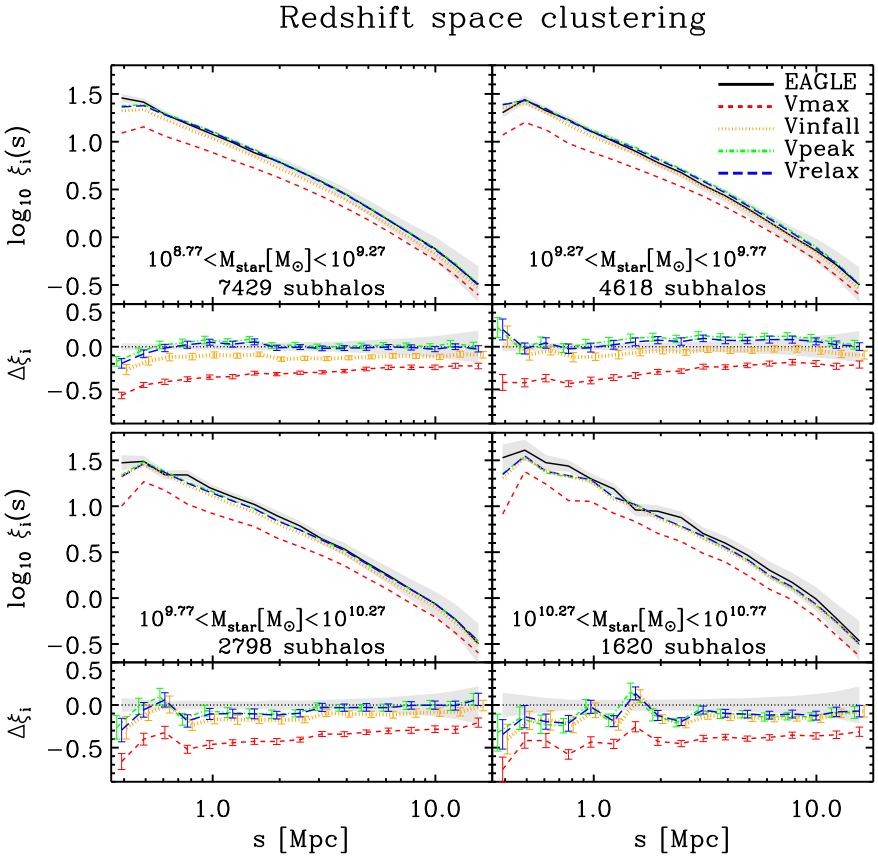


Figure 2.8: Same as Fig. 2.7 but for correlation functions computed in redshift space. The agreement between the clustering of EAGLE galaxies and $V_{\rm peak}$ and $V_{\rm relax}$ galaxies is even better in redshift space than in real space for the two lowest stellar mass bins. The main reason of the improvement on small scales is that most of the galaxies separated by those scales in redshift space are at larger distances in real space, where $V_{\rm peak}$ and $V_{\rm relax}$ galaxies accurately reproduce the clustering of EAGLE galaxies.

Redshift-space Correlation Function

Fig. 2.8 is analogous to Fig. 2.7 but for the redshift-space 2PCFs. We compute 2PCF in redshift-space because they are more directly comparable with observations than the 2PCF in real-space. We transform real- to redshift-space coordinates ($\mathbf{r}$ and $\mathbf{s}$, respectively) in the plane-parallel approximation: $\mathbf{s} = \mathbf{r} + (1+z)(\mathbf{v}\cdot\hat{\mathbf{k}})/H(z)$, where $\mathbf{v}$ the peculiar velocity, $H(z)$ is the Hubble parameter at redshift z, and $\hat{\mathbf{k}}$ is the unit vector along the z direction. On scales greater than 6 Mpc, this transformation enhances the clustering signal due to the Kaiser effect (Kaiser 1987). On smaller scales, motions inside virialised structures produce the so-called finger-of-god effect, smoothing the correlation function.

The differences between the SHAM flavours are qualitatively similar in real and redshift space: $V_{\rm max}$ underpredicts the clustering on all scales and for all $M_{\rm star}$ bins, the remaining SHAM flavours are statistically compatible with EAGLE on scales $\gtrsim 1$ Mpc, and the clustering amplitude of $V_{\rm infall}$ is systematically below that of $V_{\rm relax}$ and $V_{\rm peak}$. On the other hand, compared with the real-space 2PCFs, there is better agreement between $V_{\rm relax}$, $V_{\rm peak}$ and EAGLE on small scales for the two lowest mass bins. This improvement is likely a result of two effects. First, a considerable fraction of close pairs in redshift space will be much further apart in real space, and hence better modelled by SHAM. Second, the incorrect HOD that SHAM galaxies show can be compensated by a stronger smoothing of the 2PCF: a greater number of satellites in high-mass haloes would increase the small-scale clustering, but these satellites would also have a higher velocity dispersion.

If the agreement between SHAM and EAGLE galaxies were reached because of the cancellation of different sources of error, then this would impact other orthogonal statistics, for instance, the strength of the so-called assembly bias (other examples are the high-order multipoles of the redshift space 2PCF). We explore this next.

Assembly bias

Assembly bias generically refers to the dependence of halo clustering on any halo property other than mass, such as formation time, concentration, or spin (see, e.g., Gao et al. 2005; Gao & White 2007). It has been robustly detected in DM simulations, but it is not clear what is the effect of assembly bias on galaxy clustering. This is because a given galaxy sample will typically be a mix of haloes of different masses and properties. Although the strength of the effect depends on the assumptions of the underlying galaxy formation model, semi-analytic galaxy formation models and SHAM both suggest that assembly bias is indeed important (Croton et al. 2007; Zentner et al. 2014; Hearin et al. 2015). To our knowledge, this issue has not yet been investigated with hydrodynamical simulations.

In this section we explore whether assembly bias is present in EAGLE and whether the different SHAM flavours are able to predict its amplitude. To quantify the effect, we will compare SHAM and EAGLE 2PCFs to those measured in shuffled galaxy catalogues, which are built following the approach of Croton et al. (2007):

1) We compute the distance between each satellite galaxy and the Centre-Of-Potential (COP) of its host halo. This distance is by definition zero for central galaxies.

2) We bin haloes according to M_{200} using a bin size of 0.04 dex. We verified that our results are independent of small changes in the bin widths.

3) We randomly shuffle the entire galaxy population between haloes in the same mass bin.

4) Finally, we assign a new position to each galaxy by moving the galaxy away from the COP of its new halo by the same distance that we calculated in 1).

Fig. 2.9 shows the mean relative difference between 100 realizations of the shuffled catalogues and the original for different bins of stellar mass. The black solid lines

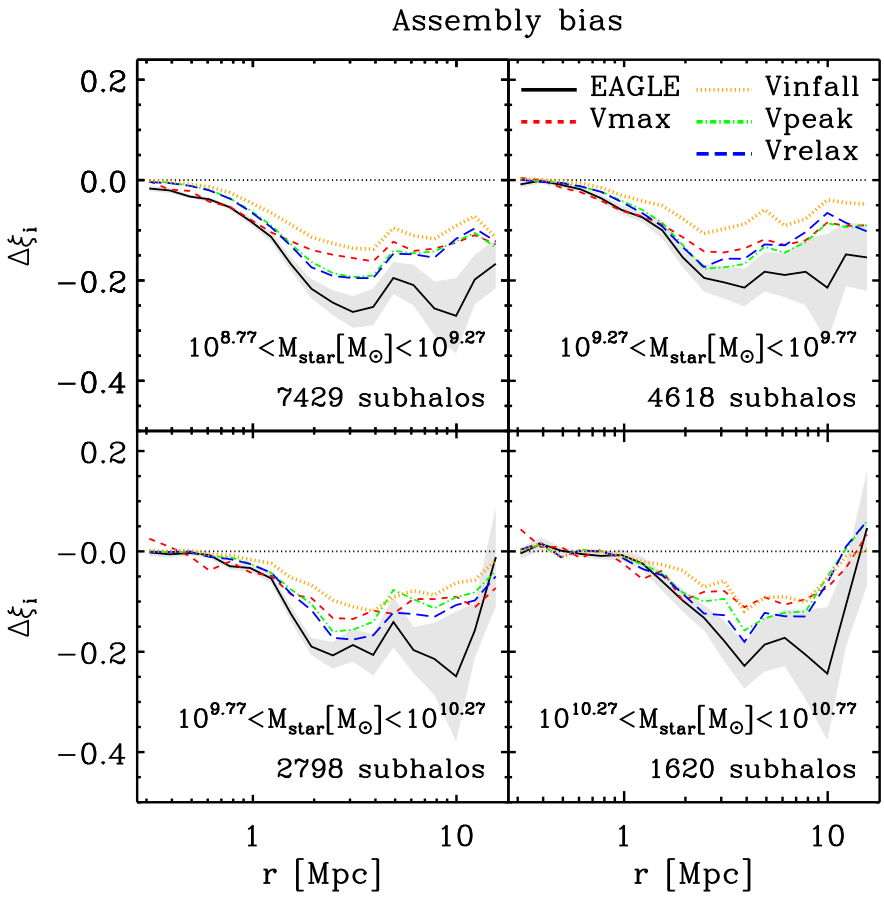


Figure 2.9: The relative difference of the 2PCFs of galaxies to that of a catalogue where galaxies are shuffled among haloes of the same mass ($\Delta\xi_i = \xi_i^{\rm shuff}/\xi_i^{\rm orig} - 1$, see §2.4.3 for more details). We adopt the same labelling as in Fig. 2.7. The grey shaded areas show the standard deviation after applying the shuffling procedure 100 times for EAGLE galaxies.

display the results for EAGLE galaxies and the coloured lines for SHAM galaxies. Since the position of galaxies/subhaloes is independent of the environment in the shuffled catalogues, their clustering should depend exclusively on the host halo mass. Therefore, any deviations from zero in Fig. 2.9 can be attributed to the assembly bias. Note that on small scales the ratio goes to zero by definition since the shuffling procedure does not alter the clustering of galaxies inside the same halo[5].

We can clearly see that all shuffled catalogues underestimate the clustering amplitude for $r \gtrsim 1$ Mpc. In the case of EAGLE galaxies, the differences are $\sim 20\%$ on scales greater than 2 Mpc, roughly independent of stellar mass. This implies that assembly bias increases the clustering amplitude expected from simple HOD analyses by about $1/0.8 = 25$ %.

[5]Note that our findings would remain nearly the same if instead we shuffled centrals and satellites separately following Zentner et al. (2014). This is because centrals and satellites with the same $M_{\rm star}$ rarely reside in the same halo (see Fig. 2.5).

For SHAM galaxies, the effect goes in the same direction but is somewhat weaker for all stellar masses (although it is more statistically significant for the lowest mass bins). This can be interpreted as SHAM lacking some environmental dependence of the relation between $M_{\rm star}$ and V_i. Likely candidates are tidal stripping of stars, and/or tidal stripping, harassment, and starvation happening before a galaxy is accreted into a larger DM halo. These effects are important because the efficiency with which a given halo creates stars will depend on the large-scale environment. We will return to these issues in the next section.

Before closing this section, it is interesting to note the particular case of $V_{\rm infall}$, which was the SHAM flavour that agreed best with the real space 2PCF of EAGLE data. The fact that the strength of the assembly bias is roughly a factor of two smaller than in EAGLE supports the idea that the previous agreement was partly coincidental. Since $V_{\rm infall}$ will be reduced near large haloes due to interactions experimented by subhaloes before being accreted, the number of satellites will decrease and the 2PCF will decrease on small scales. However, this will likely occur for the wrong haloes, which will result in a misestimated amplitude for the assembly bias.

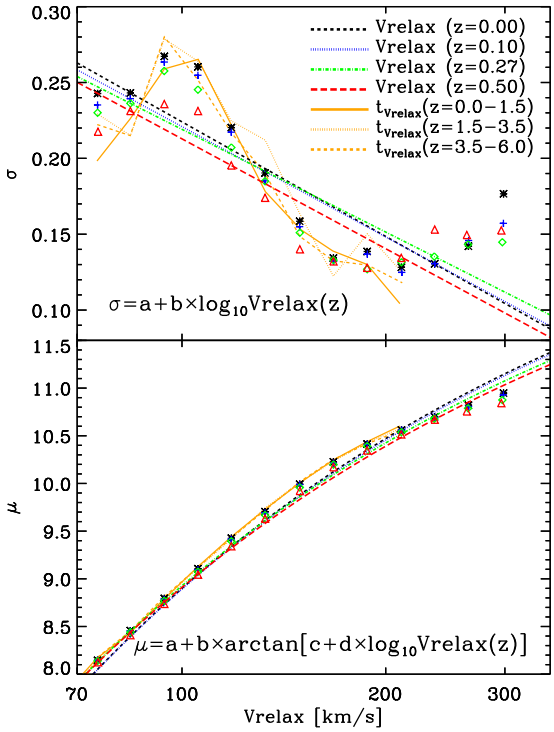


Figure 2.10: Standard deviation (top panel) and mean (bottom panel) of the Gaussian functions used to fit the dependence of the stellar mass PDF on $V_{\rm relax}$ at different redshifts. The symbols represent the measurements of the widths and the centres and the lines show the fits. Neither the scatter nor the mean of $M_{\rm star}$ and $V_{\rm relax}$ evolves significantly. The orange lines show the results for galaxies at $z = 0$ that have reached $V_{\rm relax}$ at $z = 0 - 1.5$ (solid), $z = 1.5 - 3.5$ (dotted), and $z = 3.5 - 6$ (dashed).

2.5 Testing the assumptions underlying SHAM

In the previous section we showed that SHAM reproduces the clustering of EAGLE galaxies to within 10 % on scales greater than 2 Mpc and the corresponding assembly bias reasonably well. However, small differences remain, most notably the clustering on small scales and the strength of assembly bias. In this section, we will directly test four key assumptions behind SHAM with the aim of identify the likely cause of the disagreement. Unless stated otherwise, we will employ $V_{\rm relax}$.

2.5.1 Assumption I: The relation between $M_{\rm star}$ and V_i is independent of redshift

One of the main assumptions in our implementation of SHAM is that $M_{\rm star}$ depends on the value of $V_{\rm relax}$, but *not* on the redshift at which $V_{\rm relax}$ was acquired. If this were not the case, we would expect an additional dependence on, for instance, the formation time of DM haloes. Such a redshift dependence would be particularly important for satellites, since on average they reach their value of $V_{\rm relax}$ at higher redshifts than centrals.

To test this assumption, we cross-matched the DMO and EAGLE catalogues at redshifts $z = [0, 0.1, 0.27, \text{ and } 0.5]$. We do this by assuming that the link between a pair of EAGLE-DMO structures matched at $z = 0$ carries over to their main progenitors at all higher z. Then, we construct $P(\log_{10} \mathrm{M_{star}} | \log_{10} V_i)$ at each redshift, which we fit by Gaussian functions with mean μ and standard deviation σ. In Fig. 2.10 we show the results. We can see that neither the mean nor the scatter in the relation show any strong signs of redshift dependence. Nevertheless, to estimate the impact on the clustering, we generated a new set of $V_{\rm relax}$ galaxies at $z = 0$ employing the scatter and mean derived at different redshifts. We find that the differences in the 2PCF are always below 1 %.

As a further test, we split the $z = 0$ catalogue into 3 bins according to the redshift at which $V_{\rm relax}$ was reached: $[0 - 1.5]$, $[1.5 - 3.5]$, and $[3.5 - 6]$. We overplot the mean and variance of these subsamples in Fig. 2.10 as orange lines, from which we see no obvious dependence on redshift.

Therefore, we conclude that subhaloes of a given $V_{\rm relax}$ statistically host galaxies of the same $M_{\rm star}$ at $z = 0$, independently of the time at which their $V_{\rm circ}$ reached $V_{\rm relax}$.

2.5.2 Assumption II: Baryonic physics does not affect the SHAM property of subhaloes

It is well known that baryons modify the properties of their DM hosts (Navarro et al. 1996; Gnedin & Zhao 2002; Read & Gilmore 2005; Oman et al. 2015). Notable examples are an increase in the central density of DM haloes due to adiabatic contraction, or the possible reduction due to feedback or episodic star formation events. However, SHAM assumes that the relevant property is that of the DM host in the absence of those baryonic effects.

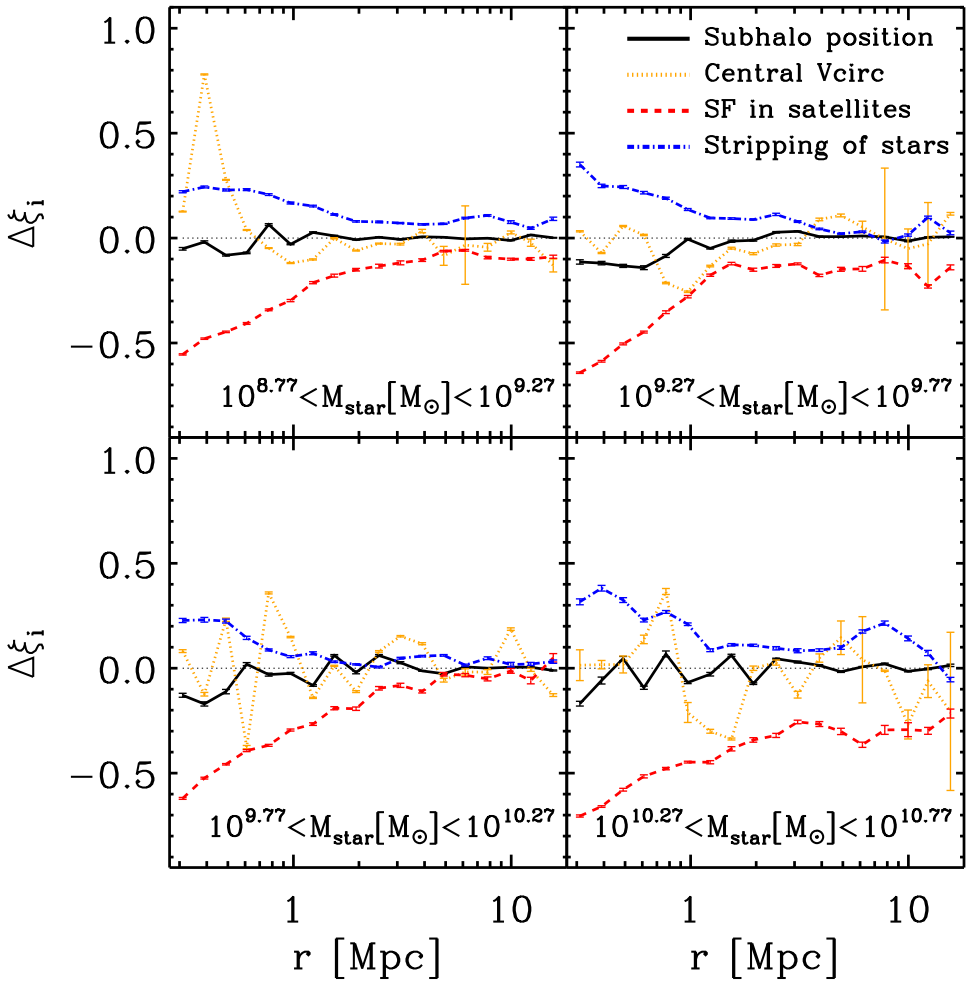


Figure 2.11: The impact on the 2PCF of different assumptions made by SHAM. Different lines compare the 2PCF of EAGLE with those of catalogues that aim to isolate different physical effects not included in SHAM in order to quantify their importance for modelling galaxy clustering. Black solid lines show the impact of baryonic effects on subhalo positions. Orange dotted lines show the impact of baryonic effects on $V_{\rm circ}$. Red dashed lines assess the importance of star formation in satellites after accretion. Blue dot-dashed lines show the impact of the stripping of stars inside massive haloes. The error bars display the jackknife statistical errors. See the main text for more details.

We estimate the impact of this assumption by comparing the 2PCFs of central galaxies in our cross-matched catalogue, which we then rank order and select using either $V_{\rm max}$ from EAGLE or $V_{\rm max}$ from their DMO counterpart. We focus on central galaxies since $V_{\rm max}$ behaves well for those objects and should be directly relevant for $V_{\rm infall}$ satellites. In addition, the cross-matched catalogue is highly complete, with less than 8% of central galaxies being excluded (see Table 2.2), thus we expect our results to be representative of the full population.

In general, we find that the values of $V_{\rm max}$ for EAGLE galaxies are ~ 5 % lower than for DMO galaxies, with a scatter of 0.08 dex. However, since the scatter is 27 % of that of $M_{\rm star}$ at a fixed $V_{\rm max}$, we expect this difference to have only a minor effect on the clustering. This is indeed what we find. The orange dotted line in Fig. 2.11 shows the relative difference of the 2PCFs. The curve is compatible with zero. Note that the noise on scales below 0.5 Mpc is caused by the small number of objects at those separations owing to the absence of satellite galaxies in this analysis.

Therefore, we conclude that baryonic effects introduce only small perturbations in V_i rank ordered catalogues and will thus only have a minor effect on SHAM predictions. In any case, the noisiness of the curves do not enable us to completely rule out small changes in the galaxy clustering due to the presence of baryons.

Table 2.6: Effect of the stripping of DM and stars from satellites, and of star formation after infall. Each value corresponds to the median of the distribution and its uncertainty computed as $\sigma = 1.4826\ \mathrm{MAD}/\sqrt{n}$, where MAD is the median absolute deviation and n the number of elements.

$M_{200}[\mathrm{M}_\odot]$	$\frac{M_{\rm DM}}{M_{\rm DM}^{\rm max}}$	$\frac{M_{\rm star}}{M_{\rm star}^{\rm max}}$	$\frac{M_{\rm star}}{M_{\rm star}^{\rm infall}}$
$M_{\rm star} = 10^{8.77} - 10^{9.27}\mathrm{M}_\odot$			
$10^{11.6} - 10^{12.6}$	0.428 ± 0.011	1.000 ± 0.000	1.714 ± 0.030
$10^{12.6} - 10^{13.6}$	0.314 ± 0.008	0.954 ± 0.002	1.828 ± 0.035
$10^{13.6} - 10^{14.6}$	0.274 ± 0.008	0.904 ± 0.004	1.446 ± 0.024
$M_{\rm star} = 10^{9.27} - 10^{9.77}\mathrm{M}_\odot$			
$10^{11.6} - 10^{12.6}$	0.458 ± 0.015	1.000 ± 0.000	1.526 ± 0.028
$10^{12.6} - 10^{13.6}$	0.329 ± 0.011	0.987 ± 0.001	1.752 ± 0.037
$10^{13.6} - 10^{14.6}$	0.278 ± 0.011	0.935 ± 0.004	1.550 ± 0.034
$M_{\rm star} = 10^{9.77} - 10^{10.27}\mathrm{M}_\odot$			
$10^{11.6} - 10^{12.6}$	0.489 ± 0.023	1.000 ± 0.000	1.360 ± 0.027
$10^{12.6} - 10^{13.6}$	0.352 ± 0.014	0.998 ± 0.000	1.532 ± 0.033
$10^{13.6} - 10^{14.6}$	0.263 ± 0.012	0.945 ± 0.004	1.433 ± 0.030
$M_{\rm star} = 10^{10.27} - 10^{10.77}\mathrm{M}_\odot$			
$10^{11.6} - 10^{12.6}$	0.670 ± 0.049	1.000 ± 0.000	1.187 ± 0.032
$10^{12.6} - 10^{13.6}$	0.386 ± 0.020	0.993 ± 0.001	1.197 ± 0.018
$10^{13.6} - 10^{14.6}$	0.238 ± 0.014	0.937 ± 0.005	1.246 ± 0.025

2.5.3 Assumption III: Baryonic physics does not affect the position of subhaloes

Another potential consequence of the presence of baryons is the modification of the positions of the subhaloes, caused by the slightly different dynamics induced by the different structure of the host halo. van Daalen et al. (2014) found this effect to be important on scales below 1 Mpc (but negligible on larger scales).

We quantify this effect by comparing the 2PCF of EAGLE galaxies in two cases; i) using their actual positions, and ii) using the position of their DMO counterparts. We show the relative difference between these two cases as a black solid line in Fig. 2.11. There are no deviations from zero on large scales and the clustering is underestimated by around 5 % on small scales. Therefore, the assumption that the presence of baryons does not modify the orbits of the subhaloes is justified for the range of scales explored here.

2.5.4 Assumption IV: For a given $V_{\rm relax}$, $M_{\rm star}$ does not depend on environment

We now address the assumption that deviations from the mean $M_{\rm star}$ at fixed $V_{\rm relax}$ are independent of the environment. Specifically, in this subsection we will investigate whether $M_{\rm star}$ at fixed $V_{\rm relax}$ is indeed uncorrelated with the host halo mass. This is a key assumption in SHAM, because it enables the modelling of galaxy clustering with a single subhalo property. Naturally, the properties of galaxies are complex functions of their merger and assembly histories, but as long as these details are not correlated with large scales, they can be treated as stochastic fluctuations within SHAM.

We start by displaying in Fig. 2.12 the median growth histories of central and satellite EAGLE galaxies within a narrow $V_{\rm relax}$ bin from 97 to 103 km s^{-1}. We show the evolution of $V_{\rm circ}$, $M_{\rm DM}$, $M_{\rm gas}$, and $M_{\rm star}$ for centrals (left panel) and satellites (right panel). Different line styles indicate the results for galaxies inside three disjoint host halo mass bins (note that the range of halo masses is different for centrals and satellites). In the case of satellites, the grey bands mark the time after these objects were accreted and brown bands mark the period after the maximum value of $M_{\rm star}(z)$ has been reached.

Interestingly, for every parameter there is a clear distinction between subhalos hosted by haloes of different masses. Central subhalos in the higher host halo mass bin formed more recently, host more massive galaxies, and have larger gas reservoirs than central subhalos hosted by less massive host haloes. Centrals hosted by haloes in the most massive bin host a galaxie with a median $M_{\rm star}$ 33 % higher than the median value for all the subhalos. On the other hand, centrals hosted by the least massive haloes have a median $M_{\rm star}$ 18 % smaller. Therefore, the difference in $M_{\rm star}$ is 0.22 dex and it corresponds to 16 % of the scatter in $M_{\rm star}$ at a fixed $V_{\rm relax}$ (c.f. Fig. 2.2), which suggests that a non-negligible fraction of the scatter can be explained by host halo variations.

The evolution of satellites is also different in distinct host halo mass bins. Subhalos that reside in more massive haloes reduce their $M_{\rm DM}$ and $V_{\rm circ}$ values more significantly, suffer from stronger stripping of gas, and stop forming stars earlier than galaxies in less

massive haloes. Furthermore, these processes appear to start prior to infall in all cases (this also serves as an example of the limitation of $V_{\rm infall}$), but the earlier the higher the halo mass (see also Behroozi et al. 2014; Bahé & McCarthy 2015). Nevertheless, and contrary to the central galaxies, the final $M_{\rm star}$ is nearly independent of the host halo mass. It is also important to mention that the median $M_{\rm star}$ for satellites is 21 % higher than for centrals, which corresponds to 0.08 dex. Thus satellite galaxies have statistically a greater $M_{\rm star}$ than central galaxies.

In general, the evolution of satellites is more complicated than that of centrals due to processes like strangulation, harassment, ram-pressure stripping, and tidal stripping (e.g., Wetzel & White 2010; Watson et al. 2012). These effects alter the growth of satellites in a non trivial way, which is not accounted for in SHAM. On the other hand, these processes are still not fully understood in detail, and it is not clear how realistically current hydrodynamical simulations like EAGLE capture them. For instance, a precise modelling of ram pressure necessarily requires a precise modelling of the intra-cluster and interstellar medium. Additionally, a precise modelling of tidal stripping requires precise morphologies of the infalling galaxies. Hence, we choose to bracket their impact on SHAM clustering predictions by considering two extreme situations.

We first consider a situation where satellite galaxies do not form or lose any stars after infall, i.e. the value of $M_{\rm star}$ is fixed at infall. The last column in Table 2.6 compares $M_{\rm star}$ at infall with $M_{\rm star}$ at $z = 0$ for galaxies hosted by haloes of different masses. The corresponding relative difference in the 2PCF is displayed by a red line in Fig. 2.11. In this case the satellites are less massive, which causes SHAM to result in a $10 - 20$ % (depending on the range of $M_{\rm star}$ considered) lower clustering signal on large scales. On small scales, the deficiency is larger, reaching more than 50 %.

The second situation we consider is one where there is no tidal stripping of stars in satellite galaxies, i.e. performing SHAM using the maximum value of $M_{\rm star}$ a galaxy has ever attained along its history, $M_{\rm star}^{\rm max}$. In Table 2.6 we compare the values of $M_{\rm star}^{\rm max}$ with $M_{\rm star}$ at $z = 0$ for different bins in stellar and host halo mass. On average, we find that the $M_{\rm star}$ reduction begins after satellites have lost about 2/3 of their $M_{\rm DM}$. We also find that this effect is stronger for low mass galaxies in higher-mass haloes, which is indeed expected due to the stronger tides. The reduction can be up to 10 % in haloes with $M_{200} > 10^{13.6} M_{\odot}$. On the other hand, this effect is essentially zero in haloes with $M_{200} < 10^{12.6} M_{\odot}$.

To quantify how the stripping of stars affects the SHAM clustering predictions, we calculate the 2PCF after selecting galaxies according to $M_{\rm star}^{\rm max}$ and compare it to our fiducial EAGLE catalogue. The result is shown by the blue dot-dashed lines in Fig. 2.11. In this case, the clustering is enhanced by about 10 % on scales greater than 1 Mpc and by up to 35 % on scales below 1 Mpc. This can be understood from the fact that the satellites are more massive, causing the satellite fraction and mean host halo mass increase, which affects the 2PCF particularly on small scales.

The two effects considered here, stellar stripping and reduced gas supply in satellites, affect the SHAM galaxy clustering to a similar magnitude but with opposite sign. In particular, for all $M_{\rm star}$ their impact is larger than the differences between SHAM and EAGLE predictions. Thus, the final galaxy clustering is sensitive to how these processes balance each other, which in turn depends sensitively on baryonic processes

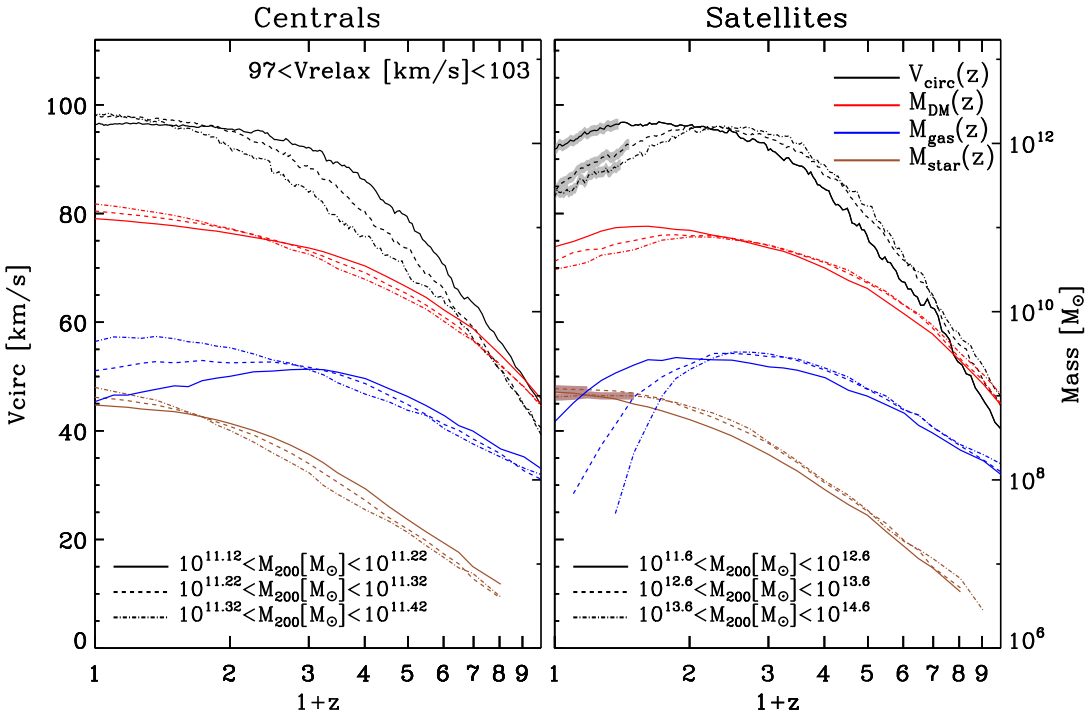


Figure 2.12: Evolution of the median of several subhalo properties along the merger history for centrals (left panel) and satellites (right panel) with $V_{\rm relax}$ between 97 and 103 km s^{-1}. The coloured lines show the evolution of the $V_{\rm circ}$, $M_{\rm DM}$, $M_{\rm gas}$, and $M_{\rm star}$, as indicated by the legend. For each component, different line styles indicate different ranges of *host* halo mass. Black lines are surrounded with a grey coloured area after $t_{\rm infall}$ and brown lines with a brown one after $t_{M_{\rm star}^{\rm max}}$. The centrals acquire $M_{\rm star}^{\rm max}$ at $z = 0$ and the ones that reside in more massive haloes end up with higher stellar masses. For satellites the behaviour of $M_{\rm star}$ is more complex. After infall, the satellites which contain gas continue forming stars until their gas is lost, but they can lose stellar mass due to tidal stripping. The subhaloes in the right panel which reside in haloes of $10^{11.6} - 10^{12.6}$, $10^{12.6} - 10^{13.6}$, $10^{13.6} - 10^{14.6}\,{\rm M}_\odot$ end up with respectively 99, 94, 91 % of their $M_{\rm star}^{\rm max}$. The stripping of DM, gas, and stars is thus more efficient for satellites in more massive host haloes (see Table 2.6).

not yet fully understood quantitatively. On the one hand, this implies an intrinsic limitation of current SHAM modelling that is reached when better than ~ 20 % accuracy is required. On the other hand, this suggests that galaxy clustering on small scales is a powerful test for the physics implemented in hydrodynamical simulations. For instance, if SHAM results were to be taken as the reality and confirmed by observations, then EAGLE would implement too weak ram-pressure stripping of massive satellite galaxies and excessive stellar stripping of low-mass galaxies in haloes with $M_{200} > 10^{12.6}\mathrm{M}_{\odot}$.

2.6 Conclusions

We have used the Ref-L100N1504 EAGLE cosmological hydrodynamical simulation to perform a detailed analysis of subhalo abundance matching for galaxies with stellar mass ranging from $10^{8.77}\mathrm{M}_{\odot}$ to $10^{10.77}\mathrm{M}_{\odot}$. We used a catalogue of paired EAGLE galaxies and subhaloes in a corresponding DM-only simulation to search for an optimal implementation of SHAM, to test its performance in terms of HOD numbers, radial number density profiles, galaxy clustering, and assembly bias, and to investigate the validity of some of the key assumptions underlying SHAM.

Our main findings can be summarized as follows:

- We argue that all current SHAM implementations use DM properties that are affected by undesired physical or numerical artefacts. Thus, we propose a new measure: V_{relax}, which is defined as the maximum circular velocity that a subhalo has reached while satisfying a relaxation criterion. We also studied SHAM using three other subhalo properties: V_{max}, the maximum circular velocity at $z = 0$; V_{infall}, the maximum circular velocity at the last time a subhalo was a central; and V_{peak}, the maximum circular velocity that a subhalo has reached. In Fig. 2.4 we show that out of the four SHAM flavours we tested, V_{relax} exhibits the strongest correlation with M_{star}, independently of the subhalo history.
- V_{infall}, V_{peak}, and V_{relax} reproduce the EAGLE predictions reasonably well (with V_{relax} performing slightly better than V_{infall} and V_{peak}):
  - Fig. 2.5 shows that the distributions of host halo masses between EAGLE and SHAM flavours match closely. In particular, the total satellite galaxy fraction agrees to within 5 %.
  - Fig. 2.7 shows that galaxy clustering strength agrees to within 10 % on scales greater than 1 Mpc and within 30 % on smaller scales. We highlight that this relation holds over four orders of magnitude in amplitude and three in length scale.
  - Fig. 2.8 shows that in redshift space the agreement improves to the point that there is no statistically significant discrepancy.
  - Assembly bias is present both in EAGLE and in its SHAM catalogues. Fig. 2.9 shows that assembly bias increases the clustering by about 20 %.

  Although small, the differences between EAGLE and SHAM are systematic and significant. We attribute these to SHAM slightly overpredicting, compared to EAGLE galaxies, the fraction of low-mass satellites in massive haloes.

- Interactions between satellites and their host haloes are very important for the amplitude of the correlation function, especially on small scales. We show in Fig. 2.11 that the difference between two extreme cases: where no stars are formed after accretion and where galaxies suffer no stripping of stars, result in differences in the amplitude of the two-point correlation function of ± 20 % on large scales and almost a factor of 2 on small scales.
- Fig. 2.12 shows that there is a relation between $M_{\rm star}$ and halo mass at fixed $V_{\rm relax}$. Centrals hosted by more massive haloes typically have higher $M_{\rm star}$, formed more recently, and contain more gas than those hosted by smaller haloes. Satellites that reside in more massive haloes typically reduce their $M_{\rm DM}$ and $V_{\rm circ}$ values more significantly, suffer from stronger stripping of gas, and stop forming stars before accretion and earlier than those in less massive haloes. The $M_{\rm star}$ of satellite galaxies at $z = 0$ is independent of the host halo mass and it is ~ 20 % greater than the $M_{\rm star}$ of central galaxies at fixed $V_{\rm relax}$.

We note that, although the box size of EAGLE (100 Mpc) is among the largest for simulations of its type, it is not large enough to ensure converged clustering properties. The lack of long wavemodes produces a few-percent excess of halos with $M \lesssim 10^{14} M_{\odot}$ and a larger deficiency of more massive halos. We expect this to reduce the satellite fraction, which may affect the shape and amplitude of overall correlation function, and might thus make our assessment of SHAM slightly too optimistic.

Overall, our results confirm the usefulness of SHAM for interpreting and modelling galaxy clustering. However, they also highlight the limits of current SHAM implementations when an accuracy better than ~ 20 % is required. Beyond this point, details of galaxy formation physics become important. For instance, SHAM assumes that the relation between $V_{\rm relax}$ and $M_{\rm star}$ is independent of the host halo mass. However, the validity of this assumption depends on how efficiently the gas content of satellite galaxies is depleted after accretion, on the importance of the stripping of stars in different environments, and on the relation between M_{DM} and $M_{\rm star}$ for centrals. EAGLE suggests that these effects depend on the host halo mass (and thus possibly on cosmological parameters), which would break the family of one-parameter SHAM models.

Fortunately, it seems possible that these physical processes can be modelled, and marginalised over, within SHAM. An interesting line of development would be the extension of SHAM to a two-parameter model, for instance a function of $V_{\rm relax}$ and $M_{\rm halo}$. This would not only reduce the systematic biases in the correlation function, but would also increase the predictive power of SHAM for centrals. We plan to explore this in the future.

Naturally, as hydrodynamical simulations improve their realism, it should be possible to better model the evolution of galaxies hosted by massive clusters, which will lead to more accurate SHAM implementations and a more accurate assessment of its performance. Ultimately, these developments will enable quick and precise predictions for the clustering of galaxies in the highly non-linear regime. In principle, this could be extended as a function of cosmology employing, e.g., cosmology-scaling methods (Angulo & White 2010; Angulo & Hilbert 2015). This opens up many interesting possibilities, such as the direct use of SHAM to optimally exploit the overwhelmingly

rich and accurate clustering measurements that are expected to arrive over the next decade.

Acknowledgements

We would like to thank Oliver Hahn and Peter Behroozi for useful discussions. Most of the parameters for EAGLE galaxies are available from the database (McAlpine+) or through interaction with the authors. This research was supported in part by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement 278594-GasAroundGalaxies, GA 267291 Cosmiway, and 321334 dustygal, the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office ([AP P7/08 CHARM]). This work used the DiRAC Data Centric system at Durham University, operated by the Institute for Computational Cosmology on behalf of the STFC DiRAC HPC Facility (www.dirac.ac.uk). This equipment was funded by BIS National E-infrastructure capital grant ST/K00042X/1, STFC capital grant ST/H008519/1, and STFC DiRAC Operations grant ST/K003267/1 and Durham University. DiRAC is part of the National E-Infrastructure. RAC is a Royal Society University Research Fellow. J.C.M acknowledges support from the "Fundación Bancaria Ibercaja" for developing this research.

Appendix A: Resolution

In this section we will present two tests that suggest that our results are not affected by the finite mass and force resolution of the EAGLE and DMO simulations. Specifically, we will explore the number of DM particles of the SHAM galaxies and compare simulations with different resolutions.

In Fig. 2.13 we show the PDF of the number of DM particles associated with central (top panel) and satellite (bottom panel) SHAM galaxies. Coloured lines show the results for different $M_{\rm star}$ bins using $V_{\rm relax}$. The detection threshold of our SUBFIND catalogues (20 particles) is marked by a vertical dashed line. The top panel shows that nearly all the central subhaloes are resolved more than 1000 DM particles. Satellites, on the other hand, are resolved with fewer particles because some of them will be lost to tidal stripping. However, since the value of $V_{\rm relax}$ will be acquired before the stripping begins, we do not expect this to affect our results. The only effect that might be important is that a subhalo can fall below the detection threshold while its counterpart galaxy is still resolved. We see that this might be the case for a very small fraction subhaloes in the lowest $M_{\rm star}$ bin. We quantify these effects next.

In Fig. 2.14 we show the number density of satellites (top panel) and the satellite fraction (bottom panel) for three different EAGLE simulations and their DMO counterparts. The black lines show the results for the same simulation used in this paper (Ref–L100N1504), the blue lines for a simulation with 25 Mpc on a side and the same resolution as Ref–L100N1504 (Ref–L025N376), and the red lines for a simulation with 25 Mpc on a side and eight times higher mass resolution than Ref–L100N1504 (Ref–L025N752). To estimate the cosmic variance, we divide Ref–L100N1504 into 64

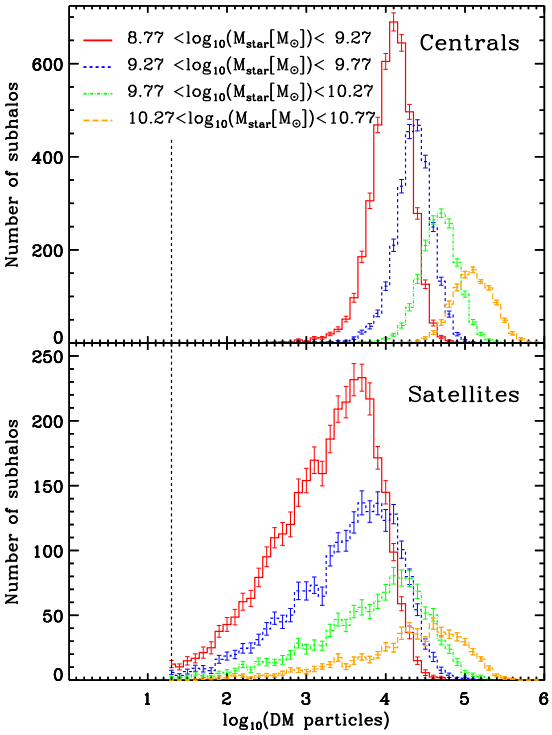


Figure 2.13: Number of DM particles in subhaloes of a given $M_{\rm star}$. The coloured lines represent the mean PDFs of 100 realizations using $V_{\rm relax}$ for different stellar mass bins and the errors are the standard deviation of the 100 realizations. The top (bottom) panel shows the PDFs of centrals (satellites). The black dashed line indicates the detection threshold of our SUBFIND catalogues. The centrals are always resolved with more than 1000 particles. However, the satellites have a tail in their distribution which reaches the detection threshold.

boxes of 25 Mpc on a side; the grey shaded areas enclose the 68 % of these boxes. The regions enclosed by vertical dotted lines in the bottom panels indicate the bins employed in §2.4.

The left two panels show that galaxies according to $M_{\rm star}$ or $V_{\rm max}$ produce almost identical satellite fractions in both (Ref–L025N752) and (Ref–L025N386), despite the former having 8 times better mass resolution. The satellite fraction coincides with our main EAGLE run for high number densities, but under-predicts the satellite fraction at low number densities. This, however, is plausibly explained by cosmic variance and the lack of long wave modes due to the smaller volume (64 times). The rightmost panel shows the DMO versions, for which the agreement between different resolutions is even better. Thus, this suggests that the results presented in this paper are not affected by the numerical resolution of our simulations.

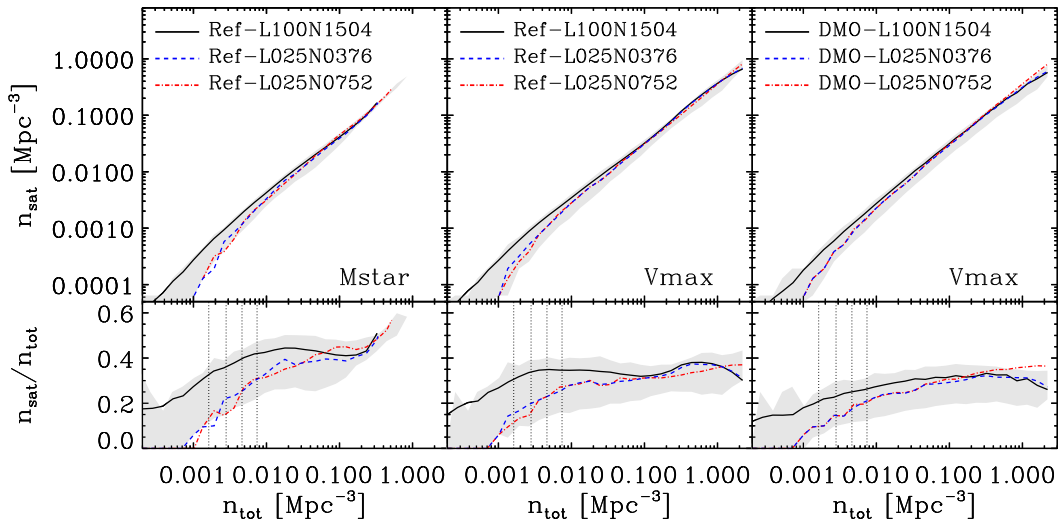


Figure 2.14: Number density of satellites (top panels) and satellite fraction (bottom panels) vs. total number density. In the left, centre, and right panels subhaloes are ordered according to $M_{\rm star}^{\rm Ref}$, $V_{\rm max}^{\rm Ref}$, and $V_{\rm max}^{\rm DMO}$ respectively. Coloured lines show the results for different simulations. The grey shaded areas enclose the 68 % of the results after dividing the simulation with the largest volume into 64 smaller boxes of 25 Mpc on a side. The regions enclosed by dotted lines indicate the bins employed in §2.4.

Appendix B: Correlation function calculation

The 2PCF counts the number of pairs at different distances in relation to the number of pairs that one would have expected from a random distribution (see, e.g., Davis et al. 1985; Peebles 2001):

$$dP = n^2[1 + \xi(\mathbf{r}_{12})]dV_1 dV_2, \tag{2.3}$$

where n is the mean density and $\xi(\mathbf{r}_{12})$ the correlation function. This equation describes the excess probability, compared with a random sample, of finding a point in an element of volume dV_2 at a distance $\mathbf{r}_{12}$ from another point in dV_1. The 2PCF is also the Fourier Transform (FT) of the power spectrum $P(\mathbf{k})$:

$$\xi(\mathbf{r}) = \frac{1}{(2\pi)^3} \int dk^3 P(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{x}}, \tag{2.4}$$

and the power spectrum is defined as:

$$\left\langle \hat{\delta}(\mathbf{k})\hat{\delta}(\mathbf{k}') \right\rangle = (2\pi)^3 \delta_{\rm D}(\mathbf{k} - \mathbf{k}')P(\mathbf{k}), \tag{2.5}$$

where $\hat{\delta}(\mathbf{k})$ is the FT of the density contrast and $\delta_D(\mathbf{k})$ is the Dirac delta function. We can use this property to quickly compute the 2PCF using Fast Fourier Transforms (FFTs). To calculate the 2PCF, we follow the following steps:

- We divide the simulation cube into 1024^3 boxes of 97.6 kpc on a side. We determine in each box the density contrast using a Cloud-In-Cell (CIC) scheme. The density contrast is defined as:

$$\delta(\mathbf{x}) = \frac{N - \langle N \rangle}{\langle N \rangle}, \tag{2.6}$$

  where N is the number of subhaloes inside one box and $\langle N \rangle$ is the total number of subhaloes in the simulation cube.

- The FT of the density field is:

$$\hat{\delta}(\mathbf{k}) = \int dx^3 e^{-i\mathbf{k}\cdot\mathbf{x}} \delta(\mathbf{x}), \tag{2.7}$$

  we compute this FT using version 3.3.3 of the Fastest Fourier Transform in the West (FFTW3; http://www.fftw.org/), a compilation of C routines for computing discrete FFTs.

- We calculate $P(\mathbf{k})$ using equation 2.5 and then we subtract the Poisson noise. The Poisson noise arises from sampling a continuous distribution with a discrete number of objects. It scales as $1/n$, where n is the number density of objects.

- The next step is to go back to real space by computing the FT of $P(\mathbf{k})$, yielding the 2PCF.

- Finally, we spherically average the correlation function obtaining the 3D 2PCF $\xi(|\mathbf{r}|)$.

By dividing the simulation cube into different number of cells, we verified that using 1024^3 boxes represents the clustering beyond 0.3 Mpc faithfully.

3 Effect of redshift errors on the galaxy clustering and the BAO

"To go wrong in one's own way is better than to go right in someone else's"

—Fyodor Dostoyevsky, *Crime and Punishment*

3.1 Introduction

A new generation of wide-field cosmological galaxy surveys will soon map the spatial distribution of hundreds of millions of galaxies over a wide range of redshifts. With these, it will be possible to characterise the expansion history of the Universe and the growth of structures with exquisite precision. Moreover, these measurements will set strong constraints on the contributors to the total energy density as a function of redshift, the law of gravity on large scales, and perhaps will offer hints to explain the accelerated expansion of the Universe (see Weinberg et al. 2013, for a review).

Some of these future galaxy surveys will employ high-resolution spectrographs, which will deliver precise estimates for the redshift of galaxies, e.g. DESI (DESI Collaboration et al. 2016), WEAVE (Dalton et al. 2014), and 4-metre Multi-Object Spectroscopic Telescope (4MOST) (de Jong 2011). Other surveys, instead, will rely on either low-resolution spectrographs, linear variable filters, or a system of narrow-band filters, such as PAUS (Martí et al. 2014), J-PAS (Benítez et al. 2014), Euclid (Laureijs et al. 2011), and Spectro-Photometer for the History of the universe, Epoch of Reionization, and ices Explorer (SPHEREx) (Doré et al. 2014, 2016). The advantage of the latter type is that they allow faster mapping speeds, provide low-resolution spectroscopic data of every pixel of the sky, and that the number of astrophysical objects for which they will produce redshifts is much greater than for the first type. Nevertheless, this approach adds non-negligible uncertainties in the measured redshifts, which will be sub-percent for surveys like J-PAS, and, in order to fully exploit their potential, their effect on cosmological observables has to be carefully studied.

The impact of photo-z errors on the galaxy clustering and on the BAO has

been explored by several authors: (Seo & Eisenstein 2003; Glazebrook & Blake 2005; Blake & Bridle 2005; Dolney et al. 2006; Seo & Eisenstein 2007; Cai et al. 2009; Benítez et al. 2009a; Sereno et al. 2015). In configuration space, adding photo-z errors to galaxy redshifts can be regarded as a smoothing operation on the galaxy field along the line-of-sight. Conversely, in Fourier space, photo-z errors reduce the amplitude of the line-of-sight modes. Despite of this, the BAO scale can still be measured, for instance, the uncertainty on the measured acoustic scale only doubles for a photo-z error of 0.3 % with respect to the no error case and for the same number density Cai et al. (2009). This has motivated surveys such as J-PAS, which aims at delivering sub-percent redshift precision for hundreds of million of galaxies employing a set of 54 contiguous FWHM ~ 150Å filters.

In this chapter we develop a complete framework for the exploitation of the BAO signal under the presence of photo-z errors. We first provide analytic expressions for how the power spectrum monopole and quadrupole (together with their variances) are affected. Then, we further study how the signal-to-noise ratio of the BAO and the cosmological information encoded in this feature depend on the photo-z error, large-scale bias, and number density of the analysed galaxy sample. We employ our findings to develop a methodology that can be applied to simulated and/or observed data to extract the BAO scale in an unbiased manner with respect to the no error case. We present our results for a wide range of number densities and photo-z errors, and for photo-z errors drawn from different probability density functions (PDFs). We also provide a fitting function that forecasts the precision with which the BAO scale can be measured and captures our numerical results accurately.

This chapter is organised as follows: in §3.2 we describe the set of cosmological simulations that we use, the way that we compute clustering statistics from them, and how we introduce photo-z errors. In §3.3 we derive analytical expressions for the impact of photo-z errors on the shape and variance of the power spectrum monopole and quadrupole. Then, in §3.4, we model how photo-z errors alter the BAO feature and the information that encodes. In §3.5 we build an unbiased model for the BAO wiggles in the power spectrum monopole and apply it to simulated samples with different photo-z errors, large-scale biases, number densities, and probability density functions. In §3.6 we explore the precision with which cosmological parameters can be measured from a survey with photo-z errors and in §3.7 we summarise our most important results.

3.2 Numerical Methods

In this section we introduce the numerical simulations that we analyse, we explain how we measure the power spectrum and its variance from these simulations, and we describe the procedure to perturb the position of the galaxies/DM particles of the simulations with photo-z errors.

3.2.1 Numerical Simulations

The first N-body calculation employed in this work is the Millennium XXL simulation (MXXL) (Angulo et al. 2012). The MXXL simulation followed 6720^3 particles of mass

$m_p = 8.456 \times 10^9 \mathrm{M}_\odot$ inside a cubical region of $3\,h^{-1}$Gpc on a side. The gravitational forces were computed with a lean version of the GAlaxies with Dark matter and Gas intEracT (GADGET) code (Springel 2005), the softening length was set to $10\,h^{-1}$ kpc, and the cosmological parameters adopted were $\Omega_m = 0.25$, $\Omega_\Lambda = 0.75$, $\Omega_b = 0.045$, $n_s = 1$, $H_0 = 73$ km s^{-1} Mpc^{-1}, and $\sigma_8 = 0.9$. Throughout this chapter we will employ a catalogue of stellar-mass selected galaxies with a number density of $n \simeq 10^{-2}\,h^3\,\mathrm{Mpc}^{-3}$ at $z = 1$, as predicted by a semi-analytic model of galaxy formation carried out on top of the MXXL merger trees. More information about this sample can be found in Angulo et al. (2014).

We complement the MXXL results with an ensemble of 300 N-body simulations, each one of them with the same volume and cosmology as the MXXL simulation but lower mass resolution. This suite has an aggregated volume of 8100 h^{-3} Gpc^3, which is large enough for statistical studies of the BAO signal. For computational efficiency, we carried out these simulations using the COmoving Lagrangian Acceleration (COLA) method (Tassev et al. 2013), where this method is able to recover the real-space power spectrum of N-body simulations to within 2 % up to $k = 0.3\,h\,\mathrm{Mpc}^{-1}$ at a fraction of the computational cost of a full N-body simulation (Howlett et al. 2015). Moreover, COLA is able to reproduce the same redshift-space power spectrum monopole and quadrupole of HOD galaxies as N-body simulations to within statistical errors up to $k = 0.2\,h\,\mathrm{Mpc}^{-1}$ (Koda et al. 2016).

Each COLA simulation evolved 1024^3 particles of mass $1.7\times10^{12}\,h^{-1}\,\mathrm{M}_\odot$ from $z = 9$ down to $z = 1$ using 10 time steps. The Gaussian initial conditions were created using 2nd order Lagrangian Perturbation theory, and the gravitational forces were computed using a Particle-Mesh algorithm with a Fourier grid of 1024^3 mesh points. Each simulation took 3 CPU hours to complete. We will not consider haloes/galaxies when analysing the COLA ensemble because their average number density is $n = 1.17 \times 10^{-6}\,h^3\,\mathrm{Mpc}^{-3}$, and at this number density the shot noise dominates the power spectrum on the scales where the BAO are located.

Together, the MXXL and the COLA suite will allow us to investigate the impact of photo-z errors on the power spectrum, its variance, and the BAO. We will only explore the $z = 1$ outputs of our simulations, which is motivated by the target redshift of future wide-field surveys. Furthermore, we assume that all the galaxies/DM particles in the simulations are at the same redshift as the simulation box.

3.2.2 Power spectrum & covariance measurements

Throughout this chapter we study the clustering in Fourier space using the power spectrum, $P(\mathbf{k})$, defined by:

$$\langle \delta(\mathbf{k})\delta(\mathbf{k}')\rangle = (2\pi)^3\delta_\mathrm{D}(\mathbf{k}-\mathbf{k}')\,P(\mathbf{k}), \tag{3.1}$$

where $\langle\,\rangle$ indicates an ensemble average, $\delta_\mathrm{D}()$ is the Dirac delta function, and $\delta(\mathbf{k})$ is the FT of the density contrast field, $\delta(\mathbf{x})$. Operationally, we compute the power spectrum by FFTs, after gridding the galaxies/particles onto a 1024^3 lattice using the CIC scheme (Hockney & Eastwood 1981). Then, we correct for the CIC window function. We expect our estimated power spectrum to be accurate to within 0.1 %

up to $k = 0.4\,h\,\mathrm{Mpc}^{-1}$ (Sefusatti et al. 2016). To estimate the anisotropic power spectrum we employ

$$\hat{P}(k,\mu) = \frac{1}{N_k}\sum_{\mathbf{k}_i} |\delta(\mathbf{k}_i)|^2, \tag{3.2}$$

where $k \equiv |\mathbf{k}|$ is the modulus of the $\mathbf{k}$ wave-vector and $\mu = \hat{\mathbf{k}}\cdot\hat{\mathbf{z}}$. This is usually named plane-parallel approximation and at $z = 1$ it only introduces a non-negligible bias in the power spectrum monopole on scales smaller than $k = 0.01\,h\,\mathrm{Mpc}^{-1}$ (Raccanelli et al. 2016). Consequently, the plane-parallel approximation do not distort the BAO at this redshift. The above sum runs over the N_k wave-vectors $\mathbf{k}_i$ that lie within a bin in (k,μ), which we define as equally spaced in $\Delta k = 0.002\,h\,\mathrm{Mpc}^{-1}$ and $\Delta\mu = 0.002$. The respective multipoles thus become

$$P_\ell(k) = \frac{2\,\ell+1}{2}\int_{-1}^{1} \mathrm{d}\mu\, \hat{P}(k,\mu)\frac{N_{k,\mu}}{N_k}\mathcal{P}_\ell(\mu), \tag{3.3}$$

where $N_{k,\mu}$ is the number of modes on the (k,μ) bin and $\mathcal{P}_\ell$ is the Legendre polynomial of order ℓ. We introduce μ dependent weights in the integral to account for the discreteness of the k-space grid and the small number of modes on large scales (Beutler et al. 2017a). Due to the finite number of μ-modes in our grid, we only trust the quadrupole ($\ell = 2$) on scales $k > 0.03\,h\,\mathrm{Mpc}^{-1}$. In addition, we apply a first order correction to remove the contribution of the shot noise to the monopole ($\ell = 0$), $P_0(k) \to P_0(k) - n^{-1}$, where n is the average number density of objects considered. Note that the shot noise vanishes for higher multipoles as it does not display an angular dependence.

An ensemble of M power spectrum measurements can be used to compute the corresponding covariance matrix:

$$\mathbf{C}_\ell(k_i,k_j) = \frac{1}{M-1}\sum_{m=1}^{M}[P_\ell^m(k_i) - \bar{P}_\ell(k_i)][P_\ell^m(k_j) - \bar{P}_\ell(k_j)], \tag{3.4}$$

where $P_\ell^m(k_i)$ is the i-th measurement of the power spectrum multipole at the scale k_i, and $\bar{P}_\ell(k_i)$ is the average value from the M simulations.

To extract the BAO scale in §3.5, we need to calculate the precision matrix, $\mathbf{C}_\ell^{-1}$. We estimate the precision matrix from the inverse of the covariance matrix, $\tilde{\mathbf{C}}_\ell^{-1}(k_i,k_j)$, using an algorithm based on a LU factorization. The precision matrix is biased when estimated from a finite number of k-bins and realizations of the power spectrum (Hartlap et al. 2007); however, it can be corrected for this as follows

$$\mathbf{C}_\ell^{-1}(k_i,k_j) = \frac{M - N_{\mathrm{bins}} - 2}{M-1}\,\tilde{\mathbf{C}}_\ell^{-1}(k_i,k_j), \tag{3.5}$$

where N_{bins} is the number of k-bins. The numerical value of the correction factor for the case of our COLA ensemble ($M = 300$) and the range of k which we use to extract the BAO scale in §3.5 is 0.60.

3.2.3 Redshift uncertainties

In our simulations, we model photo-z errors and RSD in the flat sky approximation, i.e. we perturb the comoving position of objects along the $\hat{z}$ direction, x_z:

$$x_z \to x_z + (1+z_{\rm box})\frac{v_z}{H(z_{\rm box})} + \delta(x_z) \tag{3.6}$$

where v_z is the physical peculiar velocity along the z-axis in km s^{-1}, $H(z_{\rm box})$ the Hubble parameter at the redshift of the simulation box, and $\delta(x_z)$ a random variable with PDF given by $\Pr[\delta(x_z)]$. The first term of the Right Hand Side (RHS) of the previous expression is the real-space position, the second term together with the first gives the redshift-space position, and the three terms the redshift-space position with photo-z errors.

By default, we will assume that $\Pr[\delta(x_z)]$ is a Gaussian distribution with zero mean and standard deviation $\sigma = \sigma_z(1+z_{\rm box})\,c\,H^{-1}(z_{\rm box})$, where σ_z indicates the redshift precision for samples with photo-z errors. Note that in what follows we will use σ_z and not $\sigma_z(1+z)$ to denote the redshift precision.

Photo-z errors may follow non-Gaussian PDFs in real data. For instance, the comparison of photometric and spectroscopic redshifts in the COSMOS survey showed that $\Pr[\delta(x_z)]$ is well described by a Lorentzian variate (Ilbert et al. 2009). Additionally, the $\Pr[\delta(x_z)]$ distribution for low redshift objects usually shows a tail towards higher redshifts, which is a natural consequence of imposing $z > 0$ in an otherwise symmetric PDF. Consequently, in addition to the Gaussian case, we will consider three families of functional forms for $\Pr[\delta(x_z)]$:

i) Cauchy/Lorentzian:

$$\Pr[\delta(x_z)]{\rm d}x_z = \frac{1}{\Delta\,\pi}\left[1+\left(\frac{x_z}{\Delta}\right)^2\right]^{-1}{\rm d}x_z, \tag{3.7}$$

ii) PDF1:

$$\Pr[\delta(x_z)]{\rm d}x_z = \frac{1}{2\,\Delta\,\Gamma\left(1+\frac{1}{\beta}\right)}\exp\left(-\left|\frac{x_z}{\Delta}\right|^{\beta}\right){\rm d}x_z, \tag{3.8}$$

iii) PDF2:

$$\Pr[\delta(x_z)]{\rm d}x_z = \frac{-1}{\kappa\,x_z\sqrt{2\pi}}\exp\left[-\frac{1}{2\kappa^2}\ln^2\left(-\frac{\kappa\,x_z}{\Delta}\right)\right]{\rm d}x_z, \tag{3.9}$$

where Γ is the Gamma function, Δ controls the width of the distribution, β regulates the excess kurtosis for the family PDF1, and κ directs the skewness and excess kurtosis for the family PDF2. The PDF1 distributions with $\beta < 2$ show extended wings like a Lorentzian, with $\beta = 2$ describe Gaussian distributions, and with $\beta > 2$ are boxier than a Gaussian. For the PDF2 distributions the excess kurtosis and the skewness grow with κ. Note that we disregard the possibility of interlopers, which are galaxies systematically assigned to incorrect redshifts due to misidentification of emission lines. Nevertheless, the net effect of interlopers is to increase the shot noise in the monopole.

3.3 Clustering with photometric redshift errors

In this section we derive analytical expressions for the impact of photo-z errors on the real- and redshift-space power spectrum monopole, quadrupole, and their variances, and we explain how to extend them to higher order multipoles. In all the cases, we will contrast these predictions with numerical results obtained from our set of simulations.

3.3.1 The power spectrum monopole and quadrupole

General expressions

Let us consider a set of galaxies with a real-space density contrast field $\delta_r(\mathbf{k})$ discretely sampling a Gaussian[1] field of covariance $P(\mathbf{k})$, and whose redshifts are measured through a noisy but unbiased estimator. The observed redshifts, $z_{\rm obs}$, are thus $z_{\rm obs} = z + \delta z$. Assuming that the PDF of $\delta(z)$, $\Pr[\delta(z)]$, is identical for every galaxy, we can write the redshift-space overdensity field within the Gaussian dispersion model (Kaiser 1987; Peacock & Dodds 1994):

$$\delta_z(k,\mu) = \delta_r(k)\,\mathcal{F}(k,\mu), \tag{3.10}$$

$$\mathcal{F}(k,\mu) \equiv (1+\beta\,\mu^2)\,e^{-0.5[\mu\,k\,\sigma_v(1+z)\,c/H(z)]^2}\,F(\mu k), \tag{3.11}$$

where $\beta \equiv b^{-1}\,{\rm d}\ln D(a)/{\rm d}\ln a$, b is the large-scale bias of the sample, $D(a)$ is the linear growth factor, $a = 1/(1+z)$ is the cosmological scale factor, and σ_v is a velocity dispersion induced by non-linear dynamics. The first and second terms of the RHS of Eq. 3.11 encode large- and small-scale RSD generated by the peculiar velocity of the galaxies. The last term, $F(\mu k)$, is the FT of $\Pr[\delta(z)]$. Hereafter, for brevity we will not write explicitly the dependence of $\mathcal{F}$ on μ and k.

The relation between the redshift-space power spectrum multipoles, P_ℓ, and the real-space power spectrum, $P_0^r(k)$, is

$$P_\ell(k) = (2\,\ell+1)P_0^r(k)\langle\mathcal{P}_\ell\,\mathcal{F}^2\rangle_{\hat{\mathbf{k}}}, \tag{3.12}$$

$$\langle\mathcal{P}_\ell\,\mathcal{F}^2\rangle_{\hat{\mathbf{k}}} = \frac{1}{2}\int_{-1}^{1}{\rm d}\mu\,\mathcal{P}_\ell\,\mathcal{F}^2, \tag{3.13}$$

where here and in the remainder of this chapter, $\langle ...\rangle_{\hat{\mathbf{k}}}$ brackets will denote an average over μ. For the monopole and quadrupole we get

$$P_0(k) = P_0^r(k)\,\langle\mathcal{F}^2\rangle_{\hat{\mathbf{k}}}, \tag{3.14}$$

$$P_2(k) = \frac{5}{2}\,P_0^r(k)\,\langle(3\mu^2-1)\mathcal{F}^2\rangle_{\hat{\mathbf{k}}}, \tag{3.15}$$

where it is straightforward to see that in real space ($\mathcal{F}(k,\mu) = F(\mu k)$), photo-$z$ errors create an anisotropic clustering ($P_{\ell>0} \neq 0$), even if the underlying galaxy field is isotropic. In redshift space, photo-z errors couple with the anisotropic galaxy clustering induced by RSD.

[1]Note that the Gaussian approximation neglects higher order moments due to the non-linear evolution of the density field.

The Gaussian case

In the case of a Gaussian $\Pr[\delta(z)]$, we have $F(k\mu) = \exp\left(-0.5\,\mu^2 k^2 \sigma^2\right)$, and there are analytic expressions for the terms that we need to compute the monopole and quadrupole, $\langle \mathcal{F}^2 \rangle_{\hat{\mathbf{k}}}$ and $\langle \mu^2\, \mathcal{F}^2 \rangle_{\mathbf{k}}$:

$$\langle \mathcal{F}^2 \rangle = \frac{\sqrt{\pi}}{2} \frac{\mathrm{Erf}(x)}{x} \left(1 + \frac{\beta}{x^2} + \frac{3\beta^2}{4x^4}\right) - \frac{\beta e^{-x^2}}{x^2} \left(1 + \frac{3\beta}{4x^2} \mathcal{H}_1(x)\right), \tag{3.16}$$

$$\begin{aligned} \langle \mu^2 \mathcal{F}^2 \rangle = & \frac{\sqrt{\pi}}{4} \frac{\mathrm{Erf}(x)}{x^3} \left(1 + \frac{3\beta}{x^2} + \frac{15\beta^2}{4x^4}\right) \\ & - \frac{\exp(-x^2)}{2x^2} \left(1 + \frac{3\beta}{x^2} \mathcal{H}_1(x) + \frac{15\beta^2}{4x^4} \mathcal{H}_2(x)\right), \end{aligned} \tag{3.17}$$

where $\mathcal{H}_n(x) = \sum_{i=0}^{n} \frac{2^i}{(2i+1)!!}\, x^{2i}$, !! denotes the double factorial, $x = k\,\sigma_{\rm eff}$, and $\sigma_{\rm eff} = \sqrt{\sigma_z^2 + \sigma_v^2}\,(1+z)\,c/H(z)$, i.e. the redshift uncertainties and peculiar velocities are added in quadrature (Peacock & Dodds 1994). Note that these expressions diverge as $x \to 0$, and in general $\langle \mu^n\, \mathcal{F}^m \rangle_{\mathbf{k}}$ expressions are only valid when $x > 3$. To obtain valid expressions when $x < 3$, we expand $F(\mu k)$ into a power series. The alternative expressions generated in this way are provided in Appendix A. The equivalent of Eqs. 3.16 and 3.17 in real space can be trivially obtained by setting $\beta = 0$ and $\sigma_{\rm eff} = \sigma$, where in real space and for the monopole we recover the expression provided in Peacock & Dodds (1994).

It is straightforward to see that photo-z errors suppress the amplitude of P_0 and P_2 for all wavenumbers, specially on small scales. In real space, the power spectrum is suppressed by 1.08 at $k\,\sigma = 0.5$. Moreover, photo-z errors induce a negative quadrupole, whose minimum is at $k\,\sigma \simeq 1$. In redshift space, photo-z errors couple with RSD, and thus their net impact depends on β. In general, they increase the monopole for $k\,\sigma_{\rm eff} < 1$ and decrease it for $k\,\sigma_{\rm eff} > 1$. If we consider a sample with $b = 1$ and $\beta = 1$, the monopole is increased a 66 % at $k\,\sigma_{\rm eff} = 0.5$. On small scales, the suppression is weaker than in real space, e.g. in redshift space it is 1.77 at $k\,\sigma_{\rm eff} = 2$, whereas the reduction of the monopole is 2.27 at the same scale in real space. For the quadrupole, the combined effect of photo-z and RSD is to invert its sign on scales $k\,\sigma_{\rm eff} > 1$.

Comparison with numerical simulations

In the left panel of Fig. 3.1 we display the average redshift-space power spectrum monopole of 300 COLA simulations at $z = 1$ for different photo-z errors. Note that in the case of the COLA simulations we compute the power spectrum from a sample of DM particles, where we have perturbed them following the procedure described in §3.2.3 to mimic the impact of Gaussian photo-z errors. Symbols present three cases with different photo-z errors, as indicated by the legend. As expected, photo-z errors suppress the clustering on small scales and leave the large scales unaffected. Additionally, this suppression implies that the scale at which the shot noise dominates the monopoles grows bigger with the photo-z error. The predictions of Eq. 3.16, shown by solid lines, capture the relevant effects to within 5 %. To build this model,

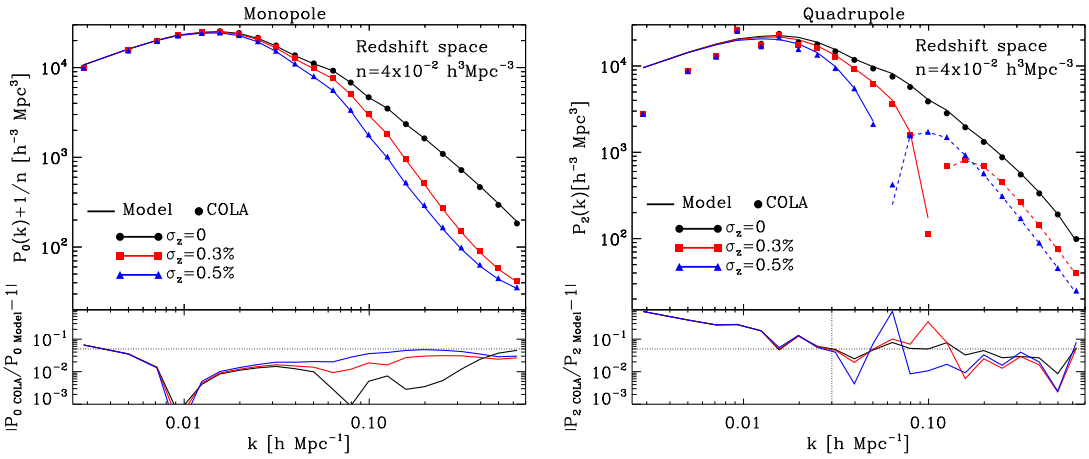


Figure 3.1: Impact of photo-z errors on the redshift-space power spectrum monopole (left panel) and quadrupole (right panel). Symbols show the average results from an ensemble of 300 N-body simulations (COLA ensemble hereafter) with different photo-z errors, whereas solid lines present our analytic predictions (Eqs. 3.14–3.16 for the monopole and Eqs. 3.15–3.17 for the quadrupole), which employ as input the real-space power spectrum and $\sigma_v = 2.94 \times 10^{-4}$. In each panel, black, red, and blue colours show the results for three Gaussian photo-z errors with a redshift precision of $\sigma_z = 0$, 0.3 %, and 0.5 %, respectively. We will employ the same colour-coding in what follows. The dashed lines of the right panel denote when the quadrupole is negative, which is for $k\,\sigma_{\rm eff} > 1$. The relative difference between the simulated data and the analytical model is shown in the bottom panels, where the horizontal dotted line indicates a 5 % discrepancy level. The discrepancies of the model and simulation data are driven by the cosmic variance on large scales, and by the inaccuracy of our RSD model on small scales. On scales smaller than $k = 0.03\,h\,{\rm Mpc}^{-1}$ and for the power spectrum quadrupole, the discrepancies are caused by the method that we employ to compute the quadrupole.

we employ the measured real-space power spectrum and we use $\sigma_v = 2.94 \times 10^{-4}$, which is obtained by fitting the average redshift-space power spectrum monopole and quadrupole of the COLA ensemble. Our model for the effect of photo-z errors on the monopole, whereas it is formally correct, it does not perfectly reproduce the data on large scales due to the cosmic variance and on small scales because of the inaccuracy of our RSD model.

In the right panel of Fig. 3.1 we show the redshift-space power spectrum quadrupole for samples with the same photo-z errors as in the left-panel. The predictions of Eqs. 3.15 and 3.17, shown by solid lines when the quadrupole is positive and by dashed lines when it is negative, capture the average results from the COLA ensemble to within 5 % on scales larger than $k = 0.03\,h\,{\rm Mpc}^{-1}$ (on smaller scales the errors are driven by the method that we employ to compute the quadrupole). Furthermore, photo-z errors cause the quadrupole to become negative on scales $k\,\sigma_{\rm eff} > 1$.

3.3.2 The variance of the monopole and quadrupole

In order to extract the BAO scale from the power spectrum, it is necessary to know the precision with which it is measured as a function of the scale. In this section, we compute the uncertainty in the power spectrum monopole and quadrupole for samples with photo-z errors.

General expressions

The effect of photo-z errors is not only to modify the shape of the power spectrum but also its variance, which is defined as the diagonal elements of the power spectrum covariance matrix, i.e:

$$\sigma^2[P] = \frac{2}{N_k} \sum_{\mathbf{k}_i} \langle |\delta(\mathbf{k}_i)|^4 \rangle - \langle \hat{P}(\mathbf{k}_i) \rangle^2, \tag{3.18}$$

where $\langle ... \rangle$ denotes the ensemble average over multiple realisations/universes. The factor two appears because only half of the modes of the power spectrum are independent due to the reality of $\delta(\mathbf{x})$. In real space, the above expression reduces to

$$\sigma^2[P] = \frac{2}{N_k} \langle P(k) \rangle^2 , \tag{3.19}$$

where this is correct for an infinite number density and in the Gaussian limit, i.e. assuming that the real and imaginary parts of $\delta(\mathbf{k})$ are Gaussian random variables with zero mean and standard deviation $P(\mathbf{k})/2$.

Combining Eqs. 3.14 and 3.18 we obtain the following expression for the variance of the redshift-space power spectrum monopole under the presence of shot noise and photo-z errors:

$$\sigma^2[P_0] = \frac{2}{N_k} \left[\frac{3}{2} \langle \mathcal{F}^4 \rangle_{\hat{\mathbf{k}}} - \frac{1}{2} \langle \mathcal{F}^2 \rangle^2_{\hat{\mathbf{k}}} + \frac{2 \langle \mathcal{F}^2 \rangle_{\hat{\mathbf{k}}}}{n P_0^r} + \frac{1}{(n P_0^r)^2} \right] (P_0^r)^2, \tag{3.20}$$

where we note that in real space and without photo-z errors ($\langle \mathcal{F}^2 \rangle_{\hat{\mathbf{k}}} = \langle \mathcal{F}^4 \rangle_{\hat{\mathbf{k}}} = 1$), our expression reduces to that provided by Colombi et al. (2009). As we mentioned earlier, we are assuming that the matter density field is a Gaussian field. However, its evolution is non-linear and it causes different k-modes to couple, which is translated into off-diagonal terms in the covariance matrix. We will study the applicability of this assumption on the scales of the BAO feature later in §3.3.2, 3.3.2, and Appendix B.

The variance of the higher order multipoles of the power spectrum do not include the last two terms in brackets in the RHS of the previous expression, as the shot noise does not have an angular dependence. Therefore, we can compute the variance as

$$\sigma^2[P_{\ell>0}] = \frac{(2\ell+1)^2}{N_k} \left[3 \langle \mathcal{P}_\ell^2 \mathcal{F}^4 \rangle_{\hat{\mathbf{k}}} - \langle \mathcal{P}_\ell \mathcal{F}^2 \rangle^2_{\hat{\mathbf{k}}} \right] (P_0^r)^2, \tag{3.21}$$

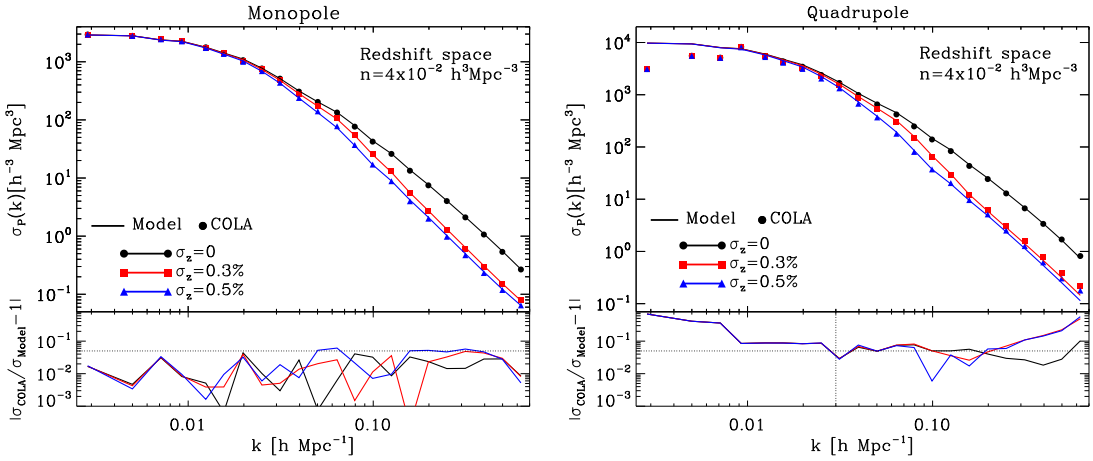


Figure 3.2: Same as Fig. 3.1 but for the standard deviation of the redshift-space power spectrum monopole (left panel) and quadrupole (right panel). In this case the numerical data is confronted with the expressions provided by Eqs. 3.20 and 3.22, respectively.

where the real-space power spectrum multipoles with $\ell > 2$ and their variances are both zero for samples without photo-z errors as $\langle \mathcal{P}_\ell^2 \mathcal{F}^4 \rangle_{\hat{\mathbf{k}}} = \langle \mathcal{P}_\ell \mathcal{F}^2 \rangle_{\hat{\mathbf{k}}} = 0$.

For the quadrupole we have

$$\sigma^2[P_2] = \frac{25}{4\,N_k} \left[3\langle (3\mu^2-1)^2 \mathcal{F}^4 \rangle_{\hat{\mathbf{k}}} - \langle (3\mu^2-1)\mathcal{F}^2 \rangle_{\hat{\mathbf{k}}}^2 \right] (P_0^r)^2 . \tag{3.22}$$

The Gaussian case

For a Gaussian Pr[$\delta(z)$], $\langle \mathcal{F}^4 \rangle_{\hat{\mathbf{k}}}$, $\langle \mu^2 \mathcal{F}^4 \rangle_{\mathbf{k}}$, and $\langle \mu^4 \mathcal{F}^4 \rangle_{\mathbf{k}}$ have analytic expressions, provided in Appendix A. Using them, it is straightforward to construct the variance of the redshift-space power spectrum monopole and quadrupole. The effect of photo-z errors is to reduce the variance of both, especially on large scales. For the monopole we have an additional effect, as the last term in brackets in the RHS of Eq. 3.20 does not depend on the photo-z error. Consequently, at a fixed scale, the shot noise contribution progressively dominates as photo-z errors increase. Because of this, the signal-to-noise of the power spectrum monopole with photo-z errors will be smaller than without them on the scales where the shot noise term dominates.

Comparison with simulations: diagonal terms

In the left and right panels of Fig. 3.2 we display the variance of the redshift-space power spectrum monopole and quadrupole for different photo-z errors, respectively. We employ the same colour coding as in Fig. 3.1. Symbols indicate again the average results from the COLA ensemble and solid lines present the prediction of Eq. 3.20 and 3.22. For the monopole, the agreement between our model and the numerical results is remarkable, showing a discrepancy always at or below the 5 % level. For the quadrupole, it is to within 10 % for $0.03\,h\,\mathrm{Mpc}^{-1} < k < 0.3\,h\,\mathrm{Mpc}^{-1}$, where on scales

smaller than $k = 0.03\,h\,\mathrm{Mpc}^{-1}$ the uncertainties are driven by the way in which we compute the quadrupole.

Comparison with simulations: off-diagonal terms

The introduction of photo-z errors in the galaxy field does not modify the covariance structure of the measured density contrast, so in particular if the real-space power spectrum covariance matrix is diagonal for samples with no photo-z errors, then so it is in redshift space with or without photo-z errors. Hence, there is no extra coupling of power spectrum modes induced by photo-z errors (or linear RSD). Nonetheless, the covariance matrix of the monopole is non-diagonal because the non-linear evolution of the matter density field couples different k-modes.

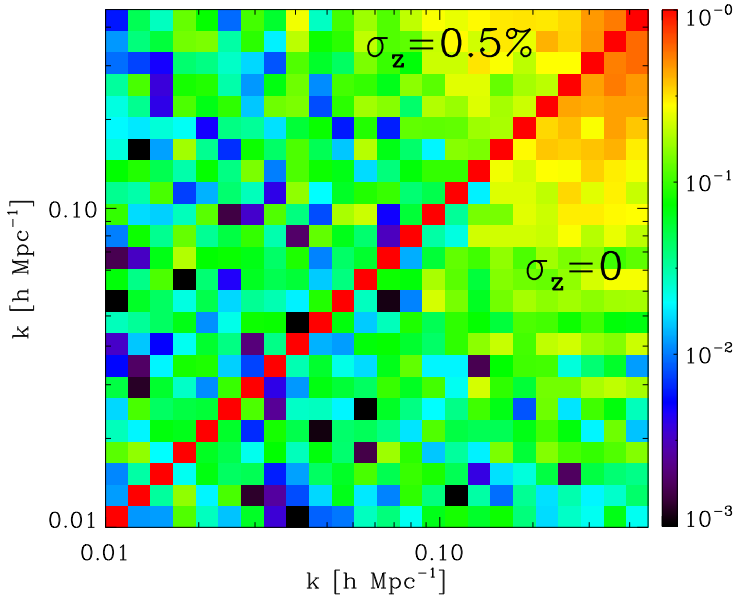


Figure 3.3: Correlation matrix of the redshift-space power spectrum monopole for $\sigma_z = 0$ (bottom triangle) and $\sigma_z = 0.5\,\%$ (top triangle) measured from the COLA ensemble. On the scales where the BAO feature is located, $k \leq 0.3\,h\,\mathrm{Mpc}^{-1}$, the off-diagonal terms are on average smaller than 0.03 in both cases. For the largest wavelengths shown, photo-z errors reduce the coupling between different k-modes.

In Fig. 3.3 we show the correlation matrix of the redshift-space power spectrum monopole $(\mathbf{C}_0(k_i,k_j)/\sqrt{\mathbf{C}_0(k_i,k_i)\mathbf{C}_0(k_j,k_j)})$ for $\sigma_z = 0$ (bottom triangle) and $\sigma_z = 0.5\,\%$ (top triangle), where these results are computed from the COLA ensemble. On large scales, the off-diagonal terms are negligible for both samples; however, on scales smaller than $k = 0.2\,h\,\mathrm{Mpc}^{-1}$ they start to be important, especially for samples with no photo-z errors. The off-diagonal terms of samples with photo-z errors are smaller than for samples without them because their effect is to decouple power spectrum modes that evolve together. Nevertheless, we can assume that the covariance matrix of the monopole is approximately diagonal on the scales where the BAO is located.

We study the effect of the off-diagonal terms of the covariance matrix when extracting the BAO scale in Appendix B.

To study whether the monopole and the quadrupole are correlated, in Fig. 3.4 we display the correlation matrices of the redshift-space power spectrum monopole ($P_0 \times P_0$), quadrupole ($P_2 \times P_2$), and their cross-correlations ($P_0 \times P_2$, $P_2 \times P_0$) for different photo-z errors. These results are computed from the COLA ensemble too. The off-diagonal terms of the correlation matrix of the quadrupole are even smaller than for the monopole, and they are negligible on the scales where the BAO are located. We also find that the monopole and quadrupole are correlated, where this correlation is positive for $k\,\sigma_{\rm eff} < 1$ and negative for $k\,\sigma_{\rm eff} > 1$. This is because the quadrupole is negative for $k\,\sigma_{\rm eff} > 1$.

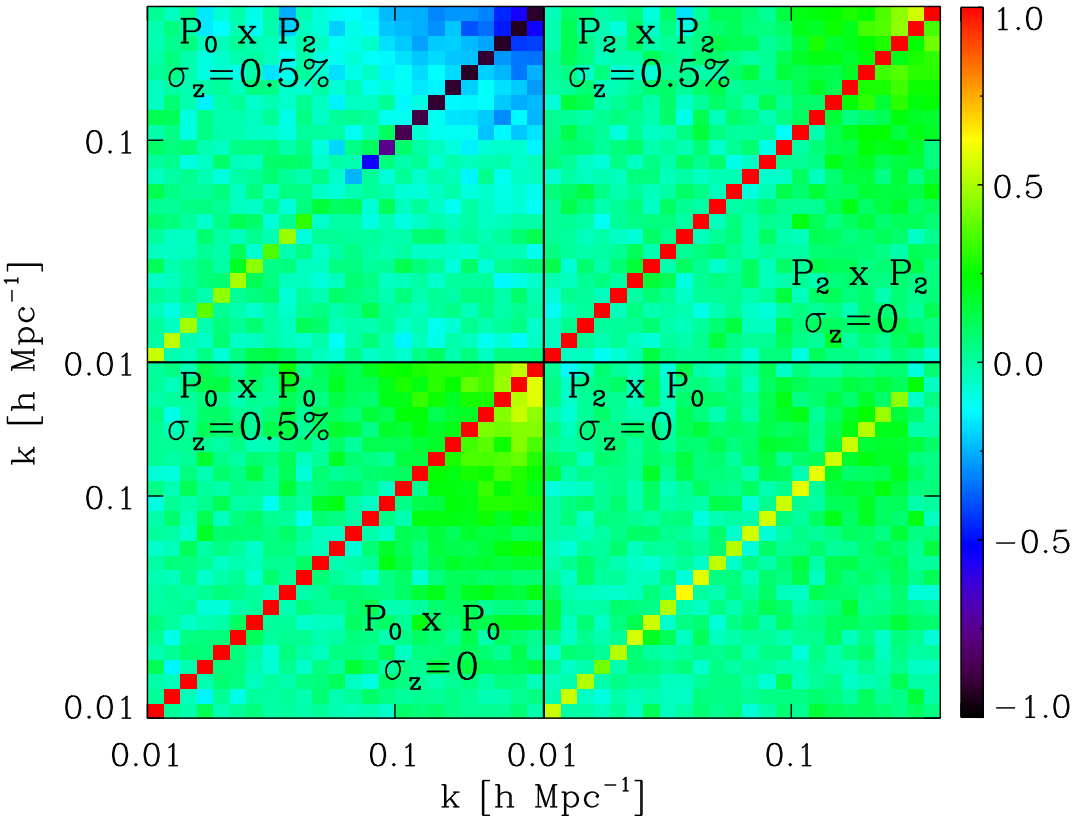


Figure 3.4: Correlation matrices of the redshift-space power spectrum monopole ($P_0 \times P_0$), quadrupole ($P_2 \times P_2$), and their cross-correlations ($P_0 \times P_2$) for $\sigma_z = 0$ and 0.5 % (top and bottom triangles). These matrices are computed from the COLA ensemble. The off-diagonal terms of the monopole and quadrupole are small on the scales shown. The monopole and quadrupole are correlated for samples with no photo-z errors whereas they are anticorrelated for samples with photo-z errors. This is because photo-z errors invert the sign of the quadrupole on scales where $k\,\sigma_{\rm eff} > 1$.

3.3.3 Signal-to-noise ratio

Let us now consider the Signal-to-Noise ratio (SNR) of the redshift-space power spectrum monopole, $P_0/\sigma[P_0]$, and quadrupole, $P_2/\sigma[P_2]$. As we mentioned before, in redshift space the effect of photo-z errors is to decrease the amplitude of the monopole, quadrupole, and their variances with respect to the no error case. Consequently, the SNR of samples with photo-z errors depends on the balance between these suppressions. Moreover, as photo-z errors invert the sign of the quadrupole on scales of the order of $k\,\sigma_{\rm eff} \simeq 1$, the SNR decreases in this case.

In §3.3.3 we introduce a toy model to understand how photo-z errors modify the SNR of the monopole, where it is straightforward to extend this model in order to study the SNR of higher order multipoles.

Toy model for the SNR of the monopole

To quantitatively understand the modifications in the SNR of the redshift-space power spectrum monopole due to photo-z errors, we build the following toy model:

$$\hat{P}_0(k) = \frac{1}{2}\hat{P}_0^r(k)\Big[\eta(\mu_1)\exp\left(-k^2\sigma_{\rm eff}^2\mu_1^2\right) + \eta(\mu_2)\exp\left(-k^2\sigma_{\rm eff}^2\mu_2^2\right)\Big], \tag{3.23}$$

where the terms in brackets provide angular contribution at only two μ-values (μ_1 and μ_2), the symbol $\hat{P}_0^r(k)$ denotes the measured real-space power spectrum, and $\eta(\mu)$ describes the contribution of large-scale RSD in a μ-bin. We will assume that $\mu_1 < \mu_2$, and thus $\eta(\mu_1) < \eta(\mu_2)$ since on linear scales $\eta(\mu)$ is a monotonically increasing function of μ.

For an ensemble average over a given k-bin we have that $\langle\hat{P}_0(k)\rangle = \frac{1}{2}P_0^r(k)[\eta(\mu_1)\exp(-k^2\,\sigma_{\rm eff}^2\,\mu_1^2) + \eta(\mu_2)\exp(-k^2\,\sigma_{\rm eff}^2\,\mu_2^2)]$, and the SNR per radial k-interval reads:

$$\mathrm{SNR} = \frac{1+\eta_{21}\exp(-k^2\,\sigma_{\rm eff}^2\,\Delta\mu^2)}{\sqrt{1+\eta_{21}^2\exp(-2\,k^2\,\sigma_{\rm eff}^2\,\Delta\mu^2)}}, \tag{3.24}$$

with $\Delta\mu^2 = \mu_2^2 - \mu_1^2$ and $\eta_{21} = \eta(\mu_2)/\eta(\mu_1)$. From this expression, we shall consider three different cases:

- No photo-z errors nor small-scale RSD, $k\,\sigma_{\rm eff} = 0$. In this case

$$\mathrm{SNR} = \frac{1+\eta_{21}}{\sqrt{1+\eta_{21}^2}}, \tag{3.25}$$

  where the SNR is always below $\sqrt{2}$, which is the value corresponding to real space.

- Very large photo-z errors, $k\,\sigma_{\rm eff} \to \infty$. In this limit, the value of the SNR is 1. It is smaller because all information is lost along the parallel modes of the power spectrum.
- Small photo-z errors, $k\,\sigma_{\rm eff} \to 0$. In this case, to first order in $(k\,\sigma_{\rm eff})^2$ we obtain

$$\mathrm{SNR} = \frac{1+\eta_{21}}{\sqrt{1+\eta_{21}^2}} + \frac{\Delta\mu^2\,\eta_{21}\,(\eta_{21}-1)}{(1+\eta_{21}^2)^{3/2}}\,(k\,\sigma_{\rm eff})^2 + \mathcal{O}\big[(k\,\sigma_{\rm eff})^4\big], \tag{3.26}$$

  where in this limit the SNR increases with respect to the case without photo-z errors nor small-scale RSD since $\eta_{21} > 1$. This behaviour must thus yield a local maximum in the SNR since for larger $k\,\sigma_{\rm eff}$ values we must recover the second case. This reflects that in this limit photo-z errors affect more the standard deviation of the power spectrum monopole than its amplitude, thus slightly increasing the SNR.

From Eq. 3.24 we find the scale corresponding to the local maximum of the SNR, $\partial(\mathrm{SNR})/\partial(k\,\sigma_{\rm eff})^2 = 0 : (k\,\sigma_{\rm eff})^2 = \ln(\eta_{21})/\Delta\mu^2$. That is, $(k\,\sigma_{\rm eff})^2 \sim 1$, as shown in Fig. 3.5.

Comparison with simulations

We now compare our analytical expressions (i.e. those derived in the previous two subsections) to the results from the COLA ensemble. In Fig. 3.5 we show the SNR of the redshift-space power spectrum monopole relative to that of the real-space power spectrum without photo-z errors. We present the results for two number densities and three different photo-z errors, as indicated by the legend. In all the cases we can see that our model, indicated by the lines, reproduces fairly well the numerical data, displayed by symbols. Independently of the value of the photo-z error, on large scales the SNR converges to the theoretical prediction for no shot noise and $\sigma_{\rm eff} = 0$, which is indicated by the horizontal dashed line. Interestingly, this limiting value is $\simeq 10\,\%$ lower than the SNR in real space, which is indicated by the horizontal dotted line. This implies that, despite the clustering enhancement due to large-scale RSD, in redshift space the SNR of the monopole is lower than in real space in the regime where shot noise is subdominant. This confirms the predictions of the toy model introduced in §3.3.3.

For the samples with photo-z errors, we appreciate an increase in the SNR relative to the case with no errors on scales where $k\,\sigma_{\rm eff} \simeq 1$, and a decrease on scales where the contribution of the shot noise is the dominant contribution in Eq. 3.20. Moreover, for $\sigma_z \lesssim 0.5\,\%$, the enhancement occurs on the scales where the BAO are located. As the BAO are suppressed by the non-linear evolution of the matter density field and non-linear RSD, this enhancement could imply that stronger cosmological constraints are derived from samples with sub-percent photo-z errors. We will return to this in the next section.

In Fig. 3.6 we display the ratio of the SNR of the redshift-space power spectrum quadrupole with photo-z errors to that with no errors. On large scales, the SNR of all the samples converges, as it happened for the monopole. On scales of the order of

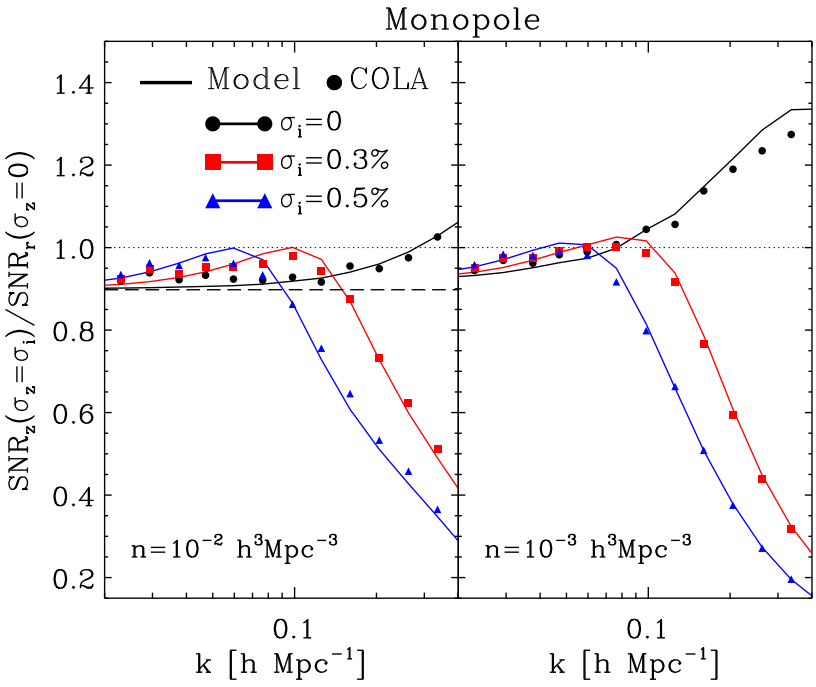


Figure 3.5: Ratio of the SNR of the redshift-space power spectrum monopole, $P_0/\sigma[P_0]$, to that of a case with no photo-z errors in real space, $P_0^r/\sigma[P_0^r]$. Symbols display this quantity computed from our COLA ensemble whereas lines do so for the analytic model of Eqs. 3.14 and 3.20. For comparison, the horizontal dashed line shows the analytic result when $n \to \infty$ and $\sigma_v = 0$, i.e. without shot noise nor small-scale RSD. On scales where $k\,\sigma_{\rm eff} \simeq 1$, the SNR of the monopole for samples with photo-z errors is greater than for samples without them. Nevertheless, on small scales it is the opposite because the scale at which the shot noise dominates the monopole grows with the photo-z error.

$k\,\sigma_{\rm eff} \simeq 1$, the SNR decreases for samples with photo-z errors. This is because photo-z errors invert the sign of the quadrupole on these scales, and thus the SNR becomes zero. However, on scales where $k\,\sigma_{\rm eff} > 1.5$, the SNR of the quadrupole with photo-z errors surpasses to that of the no error case.

3.4 Effect of photo-z errors on the BAO

We now investigate the effect of photo-z errors on the BAO and on the cosmological information that they encode. We explicitly study this for the power spectrum monopole, and we provide hints to extent our expressions to higher order multipoles.

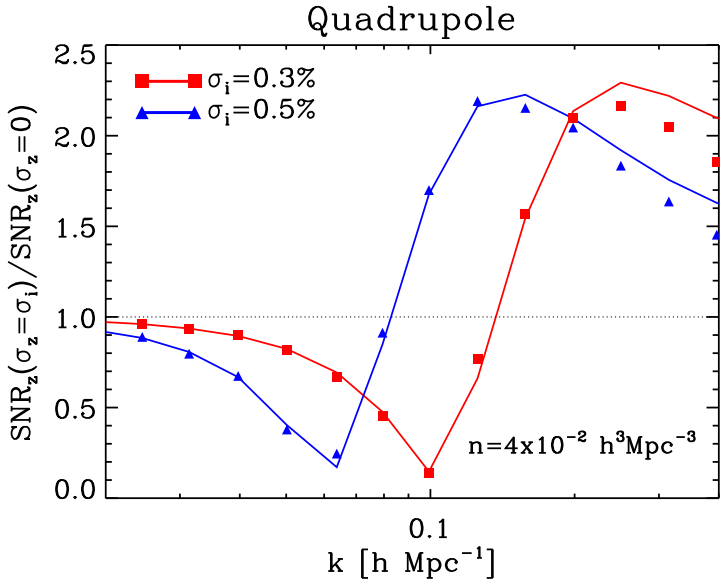


Figure 3.6: Ratio of the SNR of the redshift-space power spectrum quadrupole for samples with photo-z errors to that of samples without them. We employ the same coding as in Eq. 3.5, where the solid lines show the predictions of Eqs. 3.15 and 3.22 now. The SNR of the quadrupole for samples with photo-z errors converges to that of samples without them on large scales, it decreases on scales of the order of $k\,\sigma_{\rm eff} \simeq 1$ because the quadrupole is zero on these scales, and it is greater on scales where $k\,\sigma_{\rm eff} > 1.5$.

3.4.1 The shape of the BAO signal

Let us begin by considering the following quantity:

$$B(k) \equiv \frac{P_0(k)}{P_0^{\rm sm}(k)} - 1, \tag{3.27}$$

where $P_0^{\rm sm}$ is a no-wiggle version of the redshift-space power spectrum monopole, i.e. a power spectrum that displays the same broadband shape but no oscillatory features, i.e. without BAO. Therefore, $B(k)$ is insensitive to the overall shape of the observed power spectrum, and isolates the BAO.

Motivated by Renormalized Perturbation Theory (Crocce & Scoccimarro 2008)[2], the non-linear redshift-space power spectrum can be written as:

$$P_0(k,\mu) = [P_{0,\rm lin}(k)\,G(k,\mu) + P_{\rm mc}(k,\mu)]\;b^2\,\mathcal{F}^2, \tag{3.28}$$

where $P_{0,\rm lin}(k)$ is the linear theory power spectrum in real space, $P_{\rm mc}(k)$ is the contribution of mode coupling, and $G(k,\mu)$ is a propagator that controls the suppression

[2]Note that the smearing of the BAO signal due to non-linearities cannot be captured within the dispersion model, thus we resort to a different modelling than Eq. 3.14.

of the BAO, which is well approximated by a 2D exponential function:

$$G(k,\mu) = \exp\left\{ -\frac{1}{2}\left[(1-\mu^2)\,k^2\,\sigma_\perp^2 + \mu^2\,k^2\sigma_\parallel^2\right]\right\}, \tag{3.29}$$

where $\sigma_\parallel$ and $\sigma_\perp$ are parameters that control the loss of information due to non-linearities along and perpendicular to the line-of-sight, respectively. Note that $\sigma_\perp < \sigma_\parallel$, i.e. the BAO are more suppressed along than perpendicular to the line-of-sight, as a result of non-linear RSD (e.g. Seo & Eisenstein 2007; Sánchez et al. 2008). In redshift space the modes exactly perpendicular to the line-of-sight ($\mu = 0$) suffer the same smearing as in real space, whereas the rest of them experiment a greater suppression that grows with μ.

Let us now write a theoretical model for $B(k)$:

$$B(k) \simeq B_{\rm lin}(k)\,G_{\rm eff}(k), \tag{3.30}$$
$$G_{\rm eff}(k) = \langle G(k,\mu)\,\mathcal{F}^2\rangle_{\hat{\mathbf{k}}}\,\langle\mathcal{F}^2\rangle_{\hat{\mathbf{k}}}^{-1}, \tag{3.31}$$

where $B_{\rm lin}(k) = P_{0,\rm lin}(k)/P^{\rm sm}_{0,\rm lin}(k) - 1$, and we have assumed $P^{\rm sm}_{0,\rm lin} \approx P^{\rm sm}_{0,\rm lin}G + P^{\rm sm}_{\rm mc}$, see Crocce & Scoccimarro (2008). Therefore, the BAO wiggles in the redshift-space power spectrum monopole are suppressed by the weighted average of $G(k,\mu)$ over μ, where the weights are set by the relative decrease of line-of-sight modes caused by photo-z errors.

It is straightforward to extrapolate this model to obtain the blurring of the BAO wiggles for higher order multipoles:

$$P_\ell(k,\mu) = (2\ell+1)\left[P_{0,\rm lin}(k)\,G(k,\mu) + P_{\rm mc}(k,\mu)\right]\,b^2\,\mathcal{P}_\ell\,\mathcal{F}^2, \tag{3.32}$$
$$G_{{\rm eff},\ell}(k) = \langle G(k,\mu)\,\mathcal{P}_\ell\,\mathcal{F}^2\rangle_{\hat{\mathbf{k}}}\,\langle\mathcal{P}_\ell\,\mathcal{F}^2\rangle_{\hat{\mathbf{k}}}^{-1}. \tag{3.33}$$

The Gaussian case and comparison with simulations

In the case of Gaussian photo-z errors, Eq. 3.31 has an analytic expression:

$$\langle G(k,\mu)\,\mathcal{F}^2\rangle_{\hat{\mathbf{k}}} = e^{-\frac{1}{2}(k\,\sigma_\perp)^2}\left[\frac{\sqrt{\pi}}{2}\frac{\mathrm{Erf}(u)}{u}\left(1+\frac{\beta}{u^2}+\frac{3\beta^2}{4u^4}\right)\right.$$
$$\left. -\frac{\beta\,e^{-u^2}}{u^2}\left(1+\frac{3\beta}{4u}\mathcal{H}_1(u)\right)\right], \tag{3.34}$$

where $u = k\sqrt{\sigma^2_{\rm eff} + (\sigma^2_\parallel - \sigma^2_\perp)/2}$ and it can be employed when $u > 3$[3]. We display $G_{\rm eff}$ in the left panel of Fig. 3.7 for different photo-z errors and assuming that $\sigma_\parallel = 10\,h^{-1}\,$Mpc, $\sigma_\perp = 5\,h^{-1}\,$Mpc, $b = 1$, and the value of β expected for the cosmology of our N-body simulations. For comparison, we also show $G(k,\mu=0)$ and $G(k,\mu=1)$

[3]Note that Eq. 3.34 diverges as $u \to 0$. In Appendix A we provide a valid expression when $u < 3$.

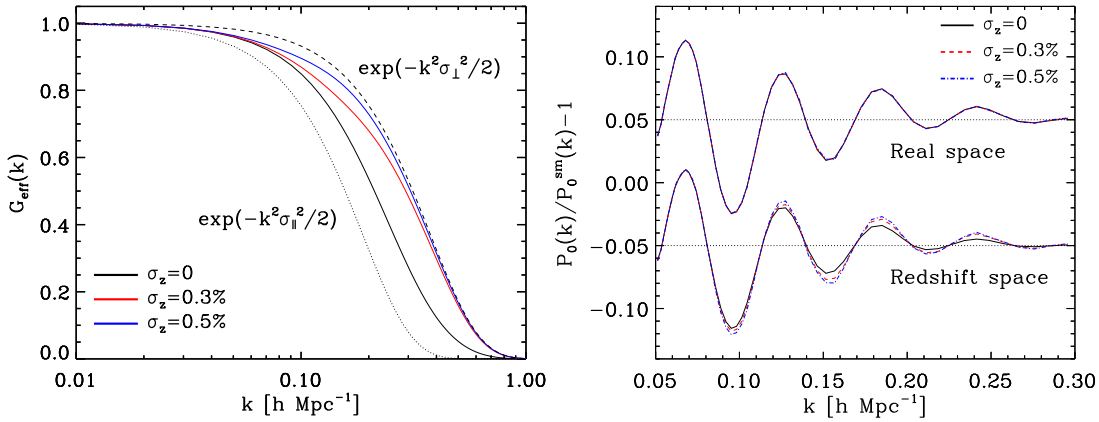


Figure 3.7: Left panel: Smoothing of the BAO wiggles due to non-linearities and RSD. The dashed and dotted lines show the suppression of the BAO feature for modes perpendicular and parallel to the line-of-sight, respectively, where the latter is stronger due to RSD. In redshift space, photo-z errors increase the weight of the k-modes perpendicular to the line-of-sight in the angular average, which make the BAO wiggles sharper. Right panel: average $B(k)$ of the COLA ensemble in real- and redshift space. We compute the no-wiggle power spectrum monopole for the COLA mocks by taking the running mean of the measured monopole.

as dashed and dotted lines, respectively. As we can see, the greater the value of the photo-z error, the smaller the contribution of $G(k, \mu = 1)$, and $G_{\rm eff} \to G(k, \mu = 0)$.

This has an interesting consequence. Since $G(k, \mu = 1) < G(k, \mu = 0)$ owing to non-linear RSD, *photo-z errors make the BAO wiggles to appear sharper in the monopole*. We explicitly show this in the right panel of Fig. 3.7, where we display the average $B(k)$ measured from the COLA ensemble in real- and redshift space. To obtain the no-wiggle power spectrum, we compute the running mean of the measured monopole from each simulation. Whereas the BAO feature is the same with and without photo-z errors in real space, in redshift space the BAO wiggles are less suppressed for the samples with photo-z errors, confirming our theoretical expectations. In real space the blurring is the same independently of the strength of photo-z errors because the reduction of the modes parallel and perpendicular to the line-of-sight is the same ($\sigma_\parallel = \sigma_\perp$), and thus

$$G_{\rm eff}(k) = e^{-\frac{1}{2}(k\,\sigma_\perp)^2}, \tag{3.35}$$

where in this case the smearing of the BAO feature just depends on the non-linear evolution of the matter density field.

3.4.2 Cosmological information on the BAO scale

We now explore how the modifications to the monopole introduced by RSD and photo-z errors affect the cosmological information encoded in the BAO feature.

Following Ross et al. (2015), let us consider a given scale in the power spectrum, $k = \sqrt{k_\parallel^2 + k_\perp^2}$. The observed scale when assuming a fiducial cosmology will be $k^{\rm fid} = \sqrt{k_\parallel^2 \alpha_\parallel^{-2} + k_\perp^2 \alpha_\perp^{-2}}$, where $\alpha_\parallel \equiv H^{\rm fid}(z)/H(z)$ and $\alpha_\perp \equiv D_A(z)/D_A^{\rm fid}(z)$. In the above expressions, $D_A(z)$ is the angular diameter distance, $H(z)$ is the Hubble parameter, and the fid superscript denotes these quantities computed in the fiducial cosmology.

The observed monopole is thus $P_0(k^{\rm fid}) = \alpha^3 P_0(k/\alpha)$ (Ballinger et al. 1996), and

$$\alpha(k) = \left\langle \mathcal{F}^2 \sqrt{\mu^2 \alpha_\parallel^2 + (1-\mu^2)\alpha_\perp^2} \right\rangle_{\hat{\mathbf{k}}}, \tag{3.36}$$

where the scale-dependence of α emerges from the scale-dependence of $\mathcal{F}$, and this expression reduces to Eq. 6 of Ross et al. (2015) for samples with no photo-z errors nor small-scale RSD. Therefore, *small-scale RSD and/or photo-z errors introduce a scale-dependence in the stretch parameter* α. Expanding the stretch parameter around the best-fitting solution to first order (Ross et al. 2015), the degeneracy is $\alpha(k) = \alpha_\parallel^{m(k)} \alpha_\perp^{n(k)}$. The values of $m(k)$ and $n(k)$ are given by:

$$m(k) \equiv \frac{1}{\langle \mathcal{F}^2 \rangle_{\hat{\mathbf{k}}}} \left. \frac{\partial \langle \alpha \rangle}{\partial \alpha_\parallel} \right|_{\alpha_\parallel = \alpha_\perp = 1} = \frac{\langle \mu^2 \mathcal{F}^2 \rangle_{\hat{\mathbf{k}}}}{\langle \mathcal{F}^2 \rangle_{\hat{\mathbf{k}}}}, \tag{3.37}$$

$$n(k) \equiv \frac{1}{\langle \mathcal{F}^2 \rangle_{\hat{\mathbf{k}}}} \left. \frac{\partial \langle \alpha \rangle}{\partial \alpha_\perp} \right|_{\alpha_\parallel = \alpha_\perp = 1} = \frac{\langle \mathcal{F}^2 \rangle_{\hat{\mathbf{k}}} - \langle \mu^2 \mathcal{F}^2 \rangle_{\hat{\mathbf{k}}}}{\langle \mathcal{F}^2 \rangle_{\hat{\mathbf{k}}}}, \tag{3.38}$$

where the higher the value of m, the more sensitive α is to the Hubble parameter. For the case of Gaussian photo-z errors, m and n have analytic expressions given by Eqs. 3.16 and 3.17.

It is straightforward to generalize these expressions to higher order multipoles. Defining $\alpha_\ell(k)$ as the stretch parameter measured from the power spectrum multipole P_ℓ, we have $\alpha_\ell(k) = \alpha_\parallel^{m_\ell(k)} \alpha_\perp^{n_\ell(k)}$, and $m_\ell(k)$ and $n_\ell(k)$ are given by

$$m_\ell(k) = \frac{\langle \mu^2 \, \mathcal{P}_\ell \, \mathcal{F}^2 \rangle_{\hat{\mathbf{k}}}}{\langle \mathcal{P}_\ell \mathcal{F}^2 \rangle_{\hat{\mathbf{k}}}}, \tag{3.39}$$

$$n_\ell(k) = 1 - m_\ell(k). \tag{3.40}$$

Going back to the monopole, the known case with $m = 1/3$ and $n = 2/3$ (Eisenstein et al. 2005) is only recovered in real space without photo-z errors. In redshift space, there is a dependence of m and n on β even if $\sigma_{\rm eff} = 0$. In general, the effect of photo-z errors and small-scale RSD is to decrease the sensitivity of the measured BAO scale on $H(z)$, whereas large-scale RSD have the opposite effect as a consequence of the line-of-sight clustering enhancement. Nonetheless, the exact degeneracy between $\alpha_\parallel$ and $\alpha_\perp$ also depends on the properties of analysed sample too, such as its number density and large-scale bias.

We can also compute the precision in measuring $\alpha(k)$ to the precision in the radial and perpendicular components using Eqs. 3.37 and 3.38:

$$\sigma^2[\alpha(k)] = m(k)^2\sigma^2[\alpha_\parallel(k)] + n(k)^2\sigma^2[\alpha_\perp(k)], \tag{3.41}$$

where the uncertainty in the radial component is $\sigma[\alpha_\parallel(k)]$ and in the perpendicular one $\sigma[\alpha_\perp(k)]$.

3.4.3 Analytical estimation of the uncertainty in α

In the previous sections we showed that the suppression of the BAO feature depends on $G_{\rm eff}$, and that the stretch parameter α is scale-dependent. In this section we use all of this to derive an analytical estimation of the uncertainty in α.

To construct an analytical estimator for the error in α, we have to take into account the precision with which the redshift-space power spectrum monopole and the BAO can be measured. Obviously, the latter depends on the former and on the smearing of the BAO wiggles. To model the amplitude of the BAO wiggles without any suppression from non-linearities nor RSD, we measure the absolute value of the local maxima and minima of $B_{\rm amp} = 1 + B_{\rm lin}$. Then, we linearly fit these values as a function of the scale, where $B_{\rm amp} = 7.82 \times 10^{-2}$ at $k = 0.05\,h\,{\rm Mpc}^{-1}$ and $B_{\rm amp} = 0.84 \times 10^{-3}$ at $k = 0.30\,h\,{\rm Mpc}^{-1}$. On the other hand, as we mentioned in §3.4.1, in redshift space the amplitude of the BAO wiggles is modulated by $G_{\rm eff}$. Therefore, to estimate the uncertainty in α we take the inverse of the amplitude of the BAO ($G_{\rm eff}\,B_{\rm amp}$) multiplied by the signal-to-noise of the monopole:

$$\hat{\sigma}[\alpha(k)] = \frac{A\,\sigma[P_0]}{P_0\,G_{\rm eff}\,B_{\rm amp}}, \tag{3.42}$$

where A is a normalization constant, and this expression is only valid on the scales where the BAO are located. The stretch parameter α is usually extracted from an interval of scales, and thus we can define an effective stretch parameter $\alpha_{\rm eff}$, with uncertainty

$$\hat{\sigma}[\alpha_{\rm eff}] = A\left(\int_{k_{\rm min}}^{k_{\rm max}} {\rm d}k\,\frac{P_0\,G_{\rm eff}\,B_{\rm amp}}{\sigma[P_0]}\right)^{-1}, \tag{3.43}$$

where we set the lower limit of the integral to $k_{\rm min} = 0.05\,h\,{\rm Mpc}^{-1}$, the upper limit to $k_{\rm max} = 0.3\,h\,{\rm Mpc}^{-1}$, and we will compute the value of the normalization constant in §3.5.3. We set these values of $k_{\rm min}$ and $k_{\rm max}$ because they are the same values that we will employ while analysing the BAO in §3.5. Note that we assume that the off-diagonal terms of the covariance matrix of the monopole are small on this interval of scales (see §3.3.2).

3.4.4 The scale-dependence of cosmological information

In §3.4.2 we showed that there is a scale dependence of the cosmological information encoded in the BAO, which introduces an additional complication while extracting information from the BAO. In the same way that we computed an effective value of the stretch parameter in the previous section, we can estimate the overall degeneracy between the parallel and perpendicular components of α when this parameter is estimated from an interval of scales. To do this, we compute the SNR weighted average value of $m(k)$ and $n(k)$ over the desired k-interval. Explicitly,

$$m_{\rm eff} = \frac{\int_{k_{\rm min}}^{k_{\rm max}} {\rm d}k \, \frac{m \, P_0 \, G_{\rm eff} \, B_{\rm amp}}{\sigma[P_0]}}{\int_{k_{\rm min}}^{k_{\rm max}} {\rm d}k \, \frac{P_0 \, G_{\rm eff} \, B_{\rm amp}}{\sigma[P_0]}}, \tag{3.44}$$

$$n_{\rm eff} = 1 - m_{\rm eff}, \tag{3.45}$$

and thus, the degeneracy between the overall precision in the parallel and perpendicular components of $\alpha_{\rm eff}$ is given by

$$\sigma^2[\alpha_{\rm eff}] = m_{\rm eff}^2 \, \sigma^2[\alpha_{{\rm eff},\parallel}] + n_{\rm eff}^2 \, \sigma^2[\alpha_{{\rm eff},\perp}]. \tag{3.46}$$

In Fig. 3.8 we display the degeneracy between the uncertainty in the parallel and perpendicular components of $\alpha_{\rm eff}$. From top to bottom, the panels display the results for $n = 10^{-2}\,h^3{\rm Mpc}^{-3}$, $10^{-3}\,h^3{\rm Mpc}^{-3}$, and $10^{-4}\,h^3{\rm Mpc}^{-3}$. The ellipses enclose the $1\,\sigma$ confidence interval on $\alpha_{\rm eff}$, and they are drawn using Eq. 3.46 with $A = 0.215$ (we compute this value in §3.5.3). In the case of no photo-z errors, the values of $m_{\rm eff}$ and $n_{\rm eff}$ are approximately the same independently of the number density of the sample. This is because there is almost no scale dependence in m and n[4]. Consequently, the ratio of the x-axis and y-axis of the ellipses, which indicates the uncertainty in $\alpha_\perp$ and $\alpha_\parallel$ respectively, is approximately identical regardless the number density.

For samples with photo-z errors, the scale dependence of m and n is much more important, where the value of m decreases with the scale. This can be inferred from the left panel of Fig. 3.7, where we can see that on small scales the suppression of the BAO for samples with large photo-z errors is approximately the same as the suppression of power spectrum modes perpendicular to the line-of-sight. Consequently, on small scales samples with photo-z errors in practice just constrain perpendicular modes, which only encode information about the angular diameter distance. This can be seen in the top-panel of Fig. 3.8, where the uncertainty in the parallel component grows with the photo-z error. As the k-value at which the contribution of the shot noise dominates the power spectrum grows with the number density, and on large scale the blurring of the BAO of samples with and without photo-z errors converges, samples with photo-z errors and small number densities will approximately have the same ratio of the x-axis and y-axis as samples with no errors. This is the consequence of cosmological information washed out on the scales where the shot noise dominates the variance of the power spectrum. In Fig. 3.8 these can be appreciated, as the shape of the ellipses is more similar for the lower number densities.

[4]We note again that there is no scale-dependence in real space; however, in redshift space there is a weak scale-dependence due to small-scale RSD.

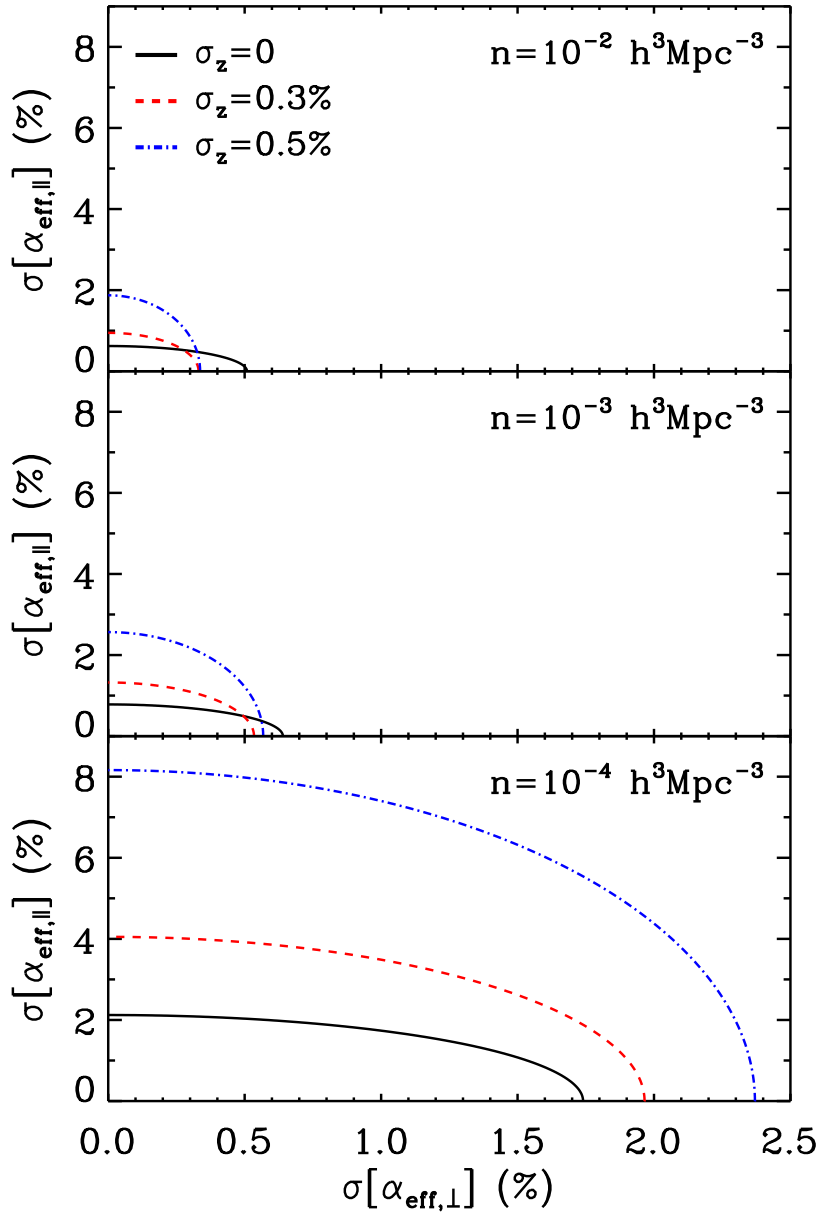


Figure 3.8: Degeneracy between the uncertainty in parallel and perpendicular components of the overall stretch parameter extracted from the redshift-space power spectrum monopole, $\alpha_{\rm eff}$. The panels show the results for different number densities and photo-z errors, as stated in the legend. The ellipses enclose the $1\,\sigma$ confidence interval for $\alpha_{\rm eff}$, where they are drawn using Eqs. 3.46 with $A = 0.215$.

Table 3.1: Degeneracy between the parallel and perpendicular component of α for samples with different number density, large-scale bias, and photo-z error. In the table we only show $m_{\rm eff}$ because $n_{\rm eff} = 1 - m_{\rm eff}$.

σ_z [%]	$n\,[h^3\,{\rm Mpc}^{-3}]$	b	$m_{\rm eff}$
0	10^{-2}	2.0	0.398
0	10^{-2}	3.5	0.371
0	10^{-3}	2.0	0.398
0	10^{-3}	3.5	0.371
0	10^{-4}	2.0	0.398
0	10^{-4}	3.5	0.371
0.3	10^{-2}	2.0	0.223
0.3	10^{-2}	3.5	0.207
0.3	10^{-3}	2.0	0.233
0.3	10^{-3}	3.5	0.210
0.3	10^{-4}	2.0	0.266
0.3	10^{-4}	3.5	0.230

In Table 3.1 we provide the value of $m_{\rm eff}$ for samples with different combinations of large-scale bias, number density, and photo-z errors. We find that the lower the large-scale bias of the sample, the stronger the constraints on the line-of-sight component of the BAO, i.e. on the Hubble parameter. This relation of the cosmological information encoded in the BAO and the characteristics of the sample highlights the need for an accurate modelling of the relevant physical and observational effects when interpreting BAO constraints in photometric galaxy surveys. In §3.5.4 we will analyse the implications of these effects on cosmological constraints derived from the BAO.

3.5 Extracting information from the BAO

In the previous sections we showed how photo-z errors modify the redshift-space power spectrum monopole, its variance, the sharpness of the BAO, and the cosmological information encoded in the BAO. In §3.5.1 we employ all this information to create a model for extracting the BAO scale from observational and/or simulated data, even under the presence of photo-z errors. We then describe our fitting procedure in §3.5.2, and in §3.5.3 we present and discuss the results of extracting the BAO feature from our simulated catalogues with different photo-z error, large-scale bias, and number density.

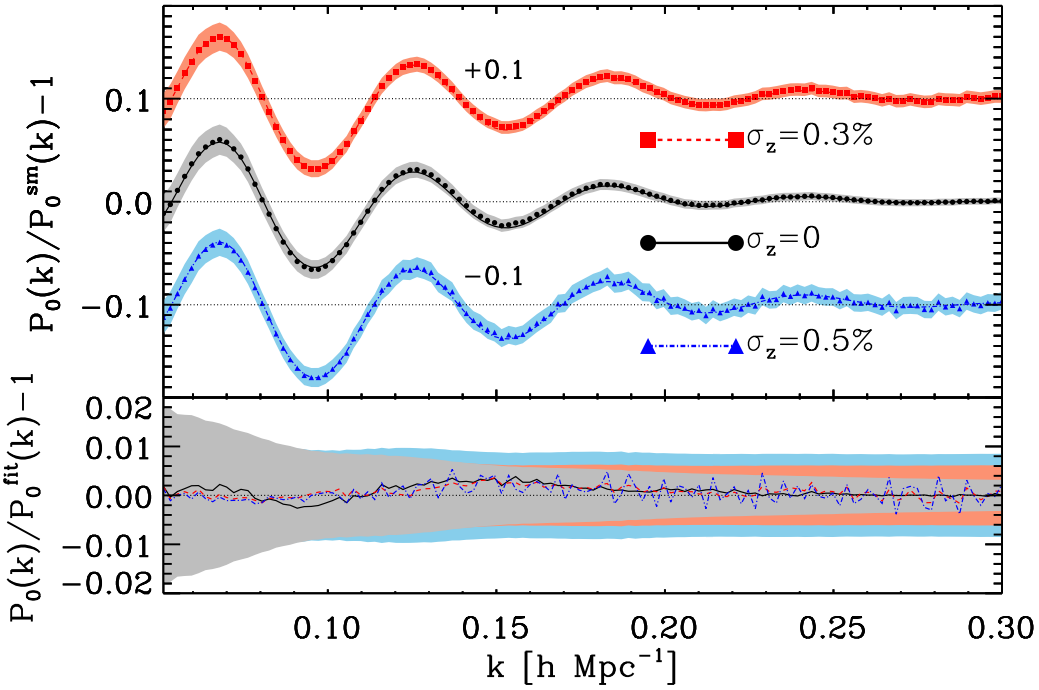


Figure 3.9: Relative difference between the redshift-space power spectrum monopole and its no-wiggle version for $n = 10^{-2}\,h^3\,\mathrm{Mpc}^{-3}$. The data points display the average monopole computed from the ensemble of COLA simulations and the lines present the average best-fit to each simulation using the model introduced by Eq. 3.47. Coloured regions enclose the area between the 84th and 16th percentiles of the scatter from mock-to-mock. The bottom panel shows the relative difference between the average results from the COLA ensemble and the average of the best-fit of our model, where its precision is to within 2 % on all the scales shown and to within 1 % for $k > 0.1\,h\,\mathrm{Mpc}^{-1}$.

3.5.1 Modelling the power spectrum monopole

Based on the expressions provided in §3.4.1, we can write the following two parameter model for the monopole under the presence of non-linearities, RSD, and photo-z errors:

$$P_T(k, \alpha_{\rm eff}, k_*) = P^{\rm sm}_{0,\rm obs}(k)\,[B_{\rm model}(k/\alpha_{\rm eff}, k_*) + 1] \tag{3.47}$$

where $P^{\rm sm}_{0,\rm obs}$ is the no-wiggle version of the observed monopole, and $B_{\rm model}$ is a function that allows suppression and dilation of the BAO wiggles, the first controlled by k_* and the second by $\alpha_{\rm eff}$, which was introduced in §3.4.3 and it is approximately equal to unity only if the length scale encoded in $B_{\rm model}(k)$ matches that of the fiducial cosmology[5].

In §3.4 we showed that the amplitude of the BAO feature is controlled by $G_{\rm eff}$, which in real space can be approximated by a Gaussian function. To keep our model

[5]The value of $\alpha_{\rm eff}$ is not exactly one because the contribution of mode coupling in Eq. 3.28 slightly shifts its value (Crocce & Scoccimarro 2008).

as simple as possible, we will approximate $G_{\rm eff}$ by a Gaussian with a width given by k_*:

$$B_{\rm model}(k,k_*) = B_{\rm lin}(k)\,\exp\left(-\frac{k^2}{2\,k_*^2}\right), \tag{3.48}$$

where k_* is a combination of the BAO smearing factors due to non-linearities, RSD, and photo-z errors.

Note that our model P_T is independent of photo-z errors (they appear just in k_*) and thus, it is similar to the template employed in the analysis of SDSS data (e.g. Percival et al. 2007, 2010; Anderson et al. 2012). Furthermore, the two free parameters that we employ, $\alpha_{\rm eff}$ and k_*, only enter in the expression for $B_{\rm model}$. Consequently, they are only constrained by BAO information, and thus our model extracts the BAO scale regardless of the overall shape of the monopole.

3.5.2 Parameter Likelihood Calculation

We assume that the probability of observing $\mathbf{d} = P_{0,\rm obs}(k)$ is given by a multivariate Gaussian distribution:

$$\Pr(\mathbf{d}|\pi) \propto \exp\left\{-\frac{1}{2}\left[\mathbf{d}-P_T(k,\pi)\right]^t\,\mathbf{C}^{-1}\left[\mathbf{d}-P_T(k,\pi)\right]\right\}, \tag{3.49}$$

where $\pi = \{\alpha_{\rm eff}, k_*\}$ are the parameters of our model P_T, which is given by Eq. 3.47. The priors on these parameters are assumed to be flat over the range: $\alpha_{\rm eff} \in [0.93, 1.07]$ and $k_* \in [0.05, 0.8]$. We note that the results do not change if we make the ranges of the priors wider. $\mathbf{C}^{-1}$ is the data precision matrix, which we compute from our COLA measurements[6] as described in §3.2.2. The range of scales considered is $k = (0.05 - 0.30)\,h\,{\rm Mpc}^{-1}$. We do not employ smaller scales since BAO wiggles are practically washed out due to non-linearities and the shot noise.

We sample the posterior probability distribution function of π employing the publicly available code EMCEE (Foreman-Mackey et al. 2013). This code is an affine invariant MCMC ensemble sampler that has been widely tested and used in multiple scientific studies. We configure the code to analyse the monopole with a chain of 100 random walkers with 5 000 steps each, and a burn-in phase of 500 steps. We check that this burn-in phase is sufficient to obtain well-behaved chains.

Additionally, we have checked that the standard deviations of the best-fit values from the COLA ensemble are compatible with the uncertainties estimated from the likelihood of each simulated catalogue.

[6] We show in Appendix B that we approximately obtain the same results in the Markov Chan Monte Carlo (MCMC) analysis when using analytical precision matrices calculated from the inverse of Eq. 3.20 as using precision matrices computed from N-body simulations.

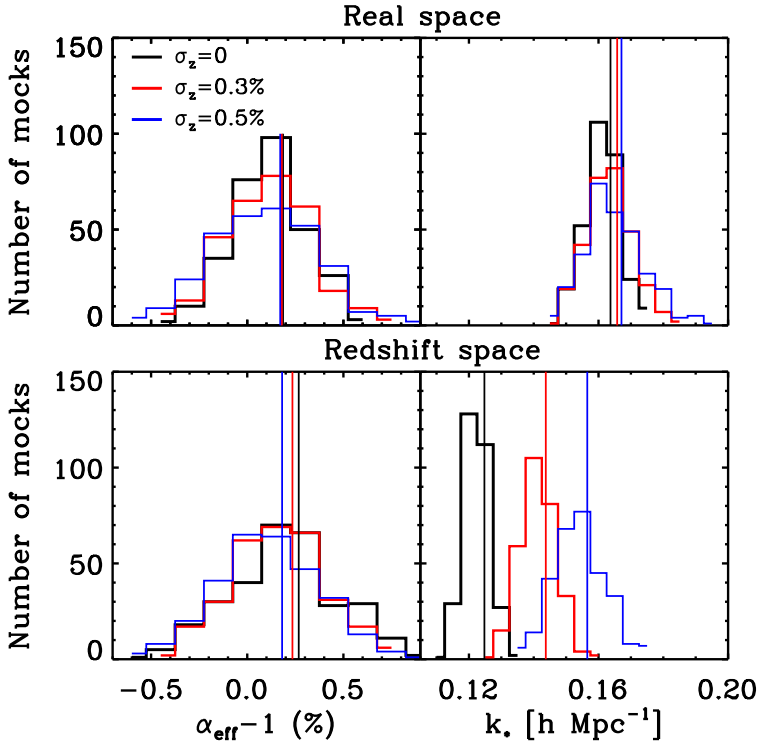


Figure 3.10: Distribution of $\alpha_{\rm eff}$ and k_* resulting from the analysis of 300 samples of DM particles with $n = 10^{-2}\,h^3\,{\rm Mpc}^{-3}$ extracted from the COLA simulations in real space (top-panels) and redshift space (bottom-panels). In real space, the value of $\alpha_{\rm eff}$ and k_* does not depend on the photo-z error. However, in redshift space $\alpha_{\rm eff}$ decreases and k_* grows with the photo-z error. Both effects are consequence of photo-z errors, the first because they reduce the coupling of different k-modes (see §3.3.2) and the second as they enhance the BAO contrast (see §3.4.1).

3.5.3 Extracting the BAO scale from the simulated catalogues

In Fig. 3.9 we show the quality of the best-fit model when applied to the COLA ensemble. The symbols indicate the average value of $B(k)$ for 300 COLA simulations, whereas the lines show the average of the best-fit of the model to each simulation. We display three cases for different photo-z errors, which have been offset for clarity. Shaded areas enclose the region between the 84th and 16th percentiles for the COLA ensemble. In all cases, the typical deviations between the data and the best-fit model are statistically insignificant, and thus our model is indeed a very good description of the measured redshift-space power spectrum monopole, which can be best seen in the bottom panel of the figure. We now explore quantitatively the results from the best-fit in a wide range of conditions.

Table 3.2: Result of the MCMC analysis for samples of DM particles with $n = 10^{-2}h^3\mathrm{Mpc}^{-3}$. We show the mean values of $\alpha_{\rm eff}$, k_*, and their uncertainties, where the error in each parameter is computed after marginalising over the other parameter. We also present the average value of the uncertainty in $\alpha_{\rm eff}$.

$\sigma_z(\%)$	$\bar{\alpha}_{\rm eff} - 1\,(\%)$	$\bar{\sigma}[\alpha_{\rm eff}]\,(\%)$	$\bar{k}_*\,[h\,\mathrm{Mpc}^{-1}]$
Real space			
0.0	0.186 ± 0.011	0.19	0.1637 ± 0.0003
0.3	0.180 ± 0.013	0.24	0.1658 ± 0.0004
0.5	0.171 ± 0.016	0.30	0.1671 ± 0.0005
Redshift space			
0	0.268 ± 0.016	0.28	0.1248 ± 0.0002
0.3	0.235 ± 0.014	0.25	0.1437 ± 0.0003
0.5	0.181 ± 0.016	0.29	0.1565 ± 0.0005

The impact of photometric redshift errors

In this section we extract the BAO scale from the power spectrum monopole of samples with different photo-z error, large-scale bias, and number density from our set of simulations. To do this, we employ the model and methodology introduced in the previous two subsections.

We start by presenting the distribution of best-fit values, $(\alpha_{\rm eff}, k_*)$, extracted from 300 independent catalogues of DM particles with number density $n = 10^{-2}\,h^3\mathrm{Mpc}^{-3}$ and different photo-z errors, extracted from the COLA ensemble. In Table 3.2 we gather these results and in Fig. 3.10 we display them, where the top panels show the best-fit values extracted from the real-space power spectrum, whereas the bottom panels do so for ones extracted from the redshift-space power spectrum. The colours indicate the magnitude of the photo-z errors, as indicated by the legend.

The top-left panel of Fig. 3.10 shows that in real space the average value of $\alpha_{\rm eff}$ is statistically compatible for samples with and without photo-z errors. This implies that *our estimator is unbiased relative to the case without photo-z errors*. Moreover, the value of the stretch parameter presents a positive shift of $\alpha_{\rm eff} - 1 \simeq 0.18\%$, which slightly decreases with σ_z and it is caused by the mode coupling induced due to the non-linear gravitational evolution of the matter density field (e.g. Angulo et al. 2008; Crocce & Scoccimarro 2008; Smith et al. 2008; Padmanabhan & White 2009). The trend with σ_z that we find is because of photo-z errors decoupling k-modes that evolve together (see §3.3.2). Moreover, this shift can in principle be corrected for using reconstruction algorithms (e.g. Eisenstein et al. 2007; Schmittfull et al. 2015) or with a recalibration of the $\alpha_{\rm eff}$ estimator. Nevertheless, it has not been proved for samples with photo-z errors.

As we can see in the bottom-left panel of Fig. 3.10, in redshift space we also find a shift in $\alpha_{\rm eff}$, which is greater than in real space. We also discover the same trend for $\alpha_{\rm eff}$ as in real space. On the other hand, for a single simulation of volume $27\,h^{-3}\,\mathrm{Gpc}^3$,

the shift in the stretch parameter is compatible with zero at the $1\,\sigma$ level in real and redshift space.

In the top-right panel of Fig. 3.10 we show the distribution of best-fit k_* values in real space, which is compatible across the catalogues with different photo-z errors. This is because in real space the blurring of the BAO feature is the same independently of the value of the photo-z error. However, as we can see in the bottom-right panel, in redshift space the value of k_* grows with the photo-z error, i.e. the larger the photo-z errors the sharper the BAO wiggles. All of this can be understood by our theoretical discussion presented in §3.4.1. In summary, photo-z errors always suppress power spectrum modes along the line-of-sight. In real space, as the BAO are isotropic, photo-z errors do not change the value of k_*, whereas in redshift space, as the BAO are more diluted along the line-of-sight due to RSD, the value of k_* grows with the photo-z error and it approaches the real-space value as bigger photo-z errors are considered.

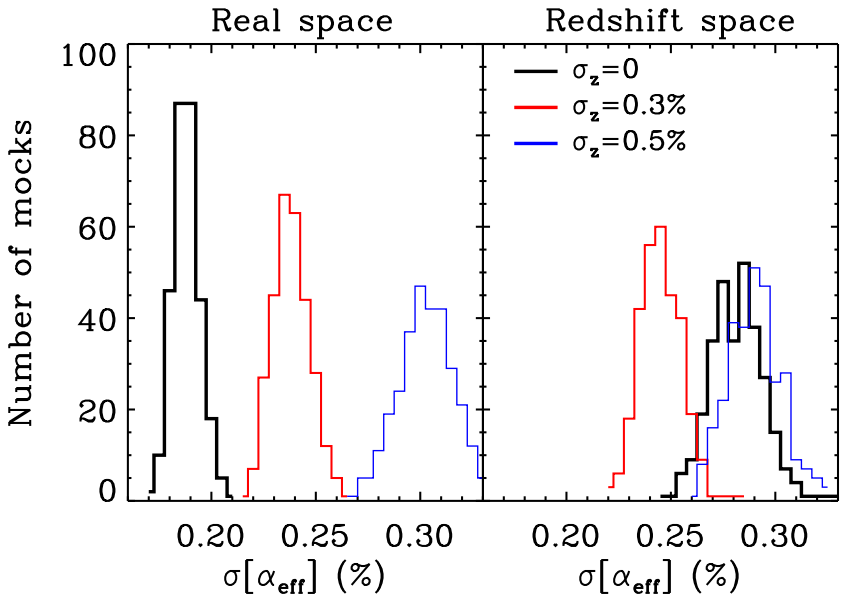


Figure 3.11: Distribution of $\sigma[\alpha_{\rm eff}]$ after marginalising over k_* for the same samples as in Fig. 3.10 in real space (left panel) and redshift space (right panel). In real space, the uncertainty in $\alpha_{\rm eff}$ grows with the photo-z error. Nonetheless, in redshift space it does not show a monotonic behaviour, and the value of $\sigma[\alpha_{\rm eff}]$ is the smallest for $\sigma_z = 0.3\,\%$. This is because samples with sub-percent photo-z errors experience a weaker smearing of the BAO and a higher SNR in the monopole on the scales where the BAO are located. On the other hand, we find that the uncertainty in $\alpha_{\rm eff}$ for $\sigma_z = 0.5\,\%$ is greater than for $\sigma_z = 0.3\,\%$ because the shot noise dominates the monopole over a larger interval of scales for the first sample, and on these scales the cosmological information is washed out.

We now pay attention to the precision with which $\alpha_{\rm eff}$ can be measured. In Fig. 3.11 we display the results for the same samples as in Fig. 3.10. In real space (left panel), the average uncertainty in $\alpha_{\rm eff}$ grows with the photo-z error. For $\sigma_z = 0.3\%$, $\alpha_{\rm eff}$ is estimated with a $\simeq 30\,\%$ less certainty than for $\sigma_z = 0$. Since the suppression

Table 3.3: Results of extracting the BAO feature from samples of the MXXL simulation in redshift space with different number densities and large-scale biases.

$\sigma_z(\%)$	b	$\alpha_{\rm eff}-1(\%)$	k_*	$\sigma[\alpha_{\rm eff}](\%)$	$\hat{\sigma}[\alpha_{\rm eff}](\%)$
$5\times10^{-3}\,h^3\,{\rm Mpc}^{-3}$					
0	1.37	0.11	0.132	0.26	0.26
0	1.58	0.17	0.128	0.26	0.26
0.3	1.37	0.24	0.157	0.24	0.25
0.3	1.58	0.29	0.153	0.24	0.24
0.5	1.37	0.23	0.167	0.29	0.29
0.5	1.58	0.27	0.167	0.28	0.29
$10^{-3}\,h^3\,{\rm Mpc}^{-3}$					
0	1.34	-0.04	0.134	0.29	0.31
0	1.84	0.10	0.125	0.29	0.28
0.3	1.34	0.17	0.168	0.28	0.32
0.3	1.84	0.16	0.151	0.27	0.28
0.5	1.34	0.16	0.189	0.35	0.39
0.5	1.84	0.10	0.173	0.32	0.34

of the BAO is independent of the photo-z error in real space, the uncertainty in $\alpha_{\rm eff}$ increases due to a larger relative contribution of the shot noise on small scales (see §3.3).

In the right-panel of Fig. 3.11 we show the uncertainty in $\alpha_{\rm eff}$ in redshift space, which is a non-monotonic function of the photo-z error. For $\sigma_z=0.3\,\%$, the precision extracting $\alpha_{\rm eff}$ increases with respect to the no error case, whereas for $\sigma_z=0.5\,\%$ it decreases. This can be understood as a balance of three effects. The first is that the SNR of the redshift-space power spectrum monopole increases at scales of the order of $k\,\sigma_{\rm eff}\simeq1$, the second that in redshift space the BAO suffer a weaker suppression as σ_z increases, and the third is that the scale at which the shot noise dominates the monopole grows with the photo-z error. Consequently, the uncertainty in $\alpha_{\rm eff}$ depends on the wavelength where $k\,\sigma_{\rm eff}\simeq1$, the value of σ_z, and the number density of the sample. For $\sigma_z=0.3\,\%$ and $n=10^{-2}\,h^3{\rm Mpc}^{-3}$, $k\,\sigma_{\rm eff}\simeq1$ occurs at $k\simeq0.1\,h\,{\rm Mpc}^{-1}$ (where the BAO are located), the smearing of the BAO is smaller than for a sample with no errors, and the shot noise does not dominate the power spectrum on the scales where the BAO are located. On the other hand, despite for the same number density and $\sigma_z=0.5\,\%$ the suppression of the BAO is even weaker, $k\,\sigma_{\rm eff}\simeq1$ occurs at $k\simeq0.05\,h\,{\rm Mpc}^{-1}$ and the shot noise starts to dominate the power spectrum on scales where the BAO are located, which implies that a considerable fraction of the BAO signal is measured with lower SNR and/or washed out.

The impact of biased tracers

In the previous section we studied the effect of photo-z errors on samples of DM particles. Here, we analyse whether the effect of photo-z errors is the same for samples of galaxies as for samples of DM particles.

The effect of the large-scale bias in real space is straightforward: it increases the amplitude of the power spectrum in all scales, and then it reduces the relative contribution of the shot noise. In redshift space, this is more complicated because the strength of the RSD depends on the large-scale bias through the parameter β. Additionally, biased tracers like galaxies are expected to display a slightly different BAO signal than DM particles (Angulo et al. 2012; Prada et al. 2016).

We study the difference between the BAO signal extracted from galaxies and DM particles by analysing samples of galaxies drawn from the MXXL simulation. We extract four samples of galaxies from this simulation according to their stellar mass: $M_*[10^{10}\,h^{-1}\,\mathrm{M}_\odot] > 2.94$, $1.37 < M_*[10^{10}\,h^{-1}\,\mathrm{M}_\odot] < 2.94$, $M_*[10^{10}\,h^{-1}\,\mathrm{M}_\odot] > 7.53$, and $1.37 < M_*[10^{10}\,h^{-1}\,\mathrm{M}_\odot] < 1.58$, where the number density of the first two is $5 \times 10^{-3}\,h^3\,\mathrm{Mpc}^{-3}$ and of the last two $10^{-3}\,h^3\,\mathrm{Mpc}^{-3}$. Then, we apply three different photo-z errors to them, and in Table 3.3 we gather the results of the MCMC analysis of these samples. We find that the bias in $\alpha_{\rm eff}$ is compatible at the $1\,\sigma$ level with zero and with the bias that we found for samples of DM matter particles in the previous section. Although the bias in the stretch parameter does not decrease with the photo-z error as it happened for DM particles, the uncertainty in $\alpha_{\rm eff}$ for a single realization is too big to make strong conclusions. On the other hand, the bias in $\alpha_{\rm eff}$ appears to decrease with the number density, as was showed in the fig. 15 of Angulo et al. (2014).

We also find that the contrast of the BAO, controlled by k_*, increases with the photo-z error and decreases with the large-scale bias, which confirms the analytical predictions of Eq. 3.31. In addition, the large-scale bias also modifies the cosmological information encoded in $\alpha_{\rm eff}$, since a higher bias decreases the dependence of the BAO scale on the Hubble parameter (see Table 3.1).

In the 5th and 6th columns of Table 3.3 we present the uncertainty in $\alpha_{\rm eff}$ computed from the MCMC analysis and the one estimated from our analytic model (Eq. 3.43), respectively. In general, the uncertainty in $\alpha_{\rm eff}$ slightly decreases with the large-scale bias. This is because a higher large-bias decreases the scale at which the monopole starts to be dominated by the shot noise. Furthermore, the uncertainty in $\alpha_{\rm eff}$ is smaller for samples with a larger number density, even if the large-scale bias is considerably smaller. Our analytic model with $A = 0.215$ is able to precisely capture all of these effects, and thus in what follows we can use this model to extent the results for DM particles to galaxies. Moreover, our model achieves the same precision for samples of DM particles as for samples of galaxies. Note that the only quantity that we employ to generate the analytic model is the average real-space power spectrum of the COLA ensemble.

The impact of the number density

In the previous sections we argued that the constraining power of the BAO wiggles depends on the scale at which the shot noise dominates the monopole. We studied

Table 3.4: Results from the MCMC analysis of the average power spectrum monopole of the COLA ensemble for different number densities and photo-z errors. We present $\alpha_{\rm eff}$ and its uncertainty after marginalising over k_*.

n	$\sigma_z(\%)$					
$[h^3\mathrm{Mpc}^{-3}]$	0.0	0.1	0.2	0.3	0.4	0.5
1×10^{-2}	0.27 ± 0.28	0.25 ± 0.27	0.25 ± 0.24	0.23 ± 0.25	0.20 ± 0.27	0.18 ± 0.29
3×10^{-3}	0.25 ± 0.30	0.24 ± 0.29	0.23 ± 0.28	0.22 ± 0.28	0.20 ± 0.31	0.18 ± 0.34
1×10^{-3}	0.27 ± 0.36	0.24 ± 0.35	0.23 ± 0.36	0.22 ± 0.38	0.18 ± 0.42	0.17 ± 0.46
8×10^{-4}	0.24 ± 0.38	0.23 ± 0.37	0.21 ± 0.38	0.19 ± 0.41	0.16 ± 0.46	0.15 ± 0.51
6×10^{-4}	0.25 ± 0.41	0.25 ± 0.41	0.23 ± 0.42	0.20 ± 0.46	0.18 ± 0.52	0.15 ± 0.59
3×10^{-4}	0.27 ± 0.53	0.26 ± 0.54	0.23 ± 0.58	0.18 ± 0.65	0.10 ± 0.72	0.04 ± 0.82
1×10^{-4}	0.30 ± 0.99	0.22 ± 1.00	0.22 ± 1.11	0.37 ± 1.28	0.20 ± 1.49	0.21 ± 1.75

this in the previous section with galaxy samples drawn from the MXXL simulation, showing that the uncertainty in the stretch parameter grows by decreasing the number density of galaxies. Moreover, we concluded that the results obtained for DM particles are approximately the same as for galaxies after accounting for a different large-scale bias. Here, we further explore the dependence of $\alpha_{\rm eff}$ on the number density using the COLA ensemble.

We apply our MCMC analysis to seven samples of DM particles extracted from the COLA simulations with different number densities: $n = [10^{-2}, 3\times10^{-3}, 10^{-3}, 8\times10^{-4}, 6\times10^{-4}, 3\times10^{-4}, 10^{-4}]\,h^3{\rm Mpc}^{-3}$, where each sample is built by randomly diluting the number of DM particles. Furthermore, we introduce photo-z errors from $\sigma_z = 0$ to $\sigma_z = 0.5\,\%$ in steps of $\sigma_z = 0.1\,\%$ to each sample, generating a total of 42 samples for each COLA simulation. In Table 3.4 we present the average results for the COLA ensemble, where we see that bias in $\alpha_{\rm eff}$ is approximately independent of the number density for samples with no photo-z errors. However, we found in the previous section that it decreased with the number density for galaxy samples. This may be consequence of using a single simulation or because of differences between diluting a sample of DM particles and galaxies.

For samples with photo-z errors, we find that the bias in $\alpha_{\rm eff}$ decreases by increasing their value. As we mentioned in §3.5.3, this is likely due to a smaller mode coupling for samples with photo-z errors. Moreover, the bias in $\alpha_{\rm eff}$ also shrinks with the number density for these samples. This is because for the same redshift error, the k-value at which the shot-noise dominates the power spectrum grows by reducing the number density. Consequently, for samples with small number densities the constraints in $\alpha_{\rm eff}$ come from larger scales than for samples with large number densities, and thus from scales are less affected by non-linearities.

In the left panel of Fig. 3.12 we display the uncertainty in $\alpha_{\rm eff}$ as a function of the number density and the photo-z error. At high number densities, the relation between $\sigma[\alpha_{\rm eff}]$ and σ_z is non-monotonic, as discussed in §3.5.3. However, as we consider lower number densities, the uncertainty in $\alpha_{\rm eff}$ starts to monotonically grow with σ_z. This is because the weaker suppression of the BAO and the higher SNR on large scales do not compensate that the interval of scales dominated by shot noise is smaller for samples without photo-z errors than for samples with them. At the typical number densities of spectroscopic galaxy surveys, e.g. $n = 3\times10^{-4}\,h^3\,{\rm Mpc}^{-3}$ for SDSS-III, the uncertainty in $\alpha_{\rm eff}$ increases monotonically with the photo-z errors, $\simeq 20\,\%$ and $\simeq 45\%$ larger for $\sigma_z = 0.3\%$ and $\sigma_z = 0.5\%$, respectively. Conversely, to reach the same accuracy on $\alpha_{\rm eff}$ as SDSS-III, the number density should be 1.5 (2.5) times larger for a photometric galaxy survey with $\sigma_z = 0.3\,\%$ ($\sigma_z = 0.5\,\%$).

In the central and right panels of Fig. 3.12 we present the uncertainty in the perpendicular and parallel components of $\alpha_{\rm eff}$, respectively, which are proportional to the precision measuring the the angular diameter distance and Hubble parameter. Note that for constructing this figure we have employed Eq. 3.44, where we have adopted the effective values of m and n corresponding to each case. As happened to the uncertainty in $\alpha_{\rm eff}$, the error in its components decrease with the number density for the same photo-z error. For $\alpha_{\rm eff,\perp}$ we find a non-monotonic behaviour of its error for large number densities, whereas for small number densities $\sigma[\alpha_{\rm eff,\perp}]$ grows with σ_z. This is due to the same reasons that caused a non-monotonic behaviour of $\sigma[\alpha_{\rm eff}]$ with

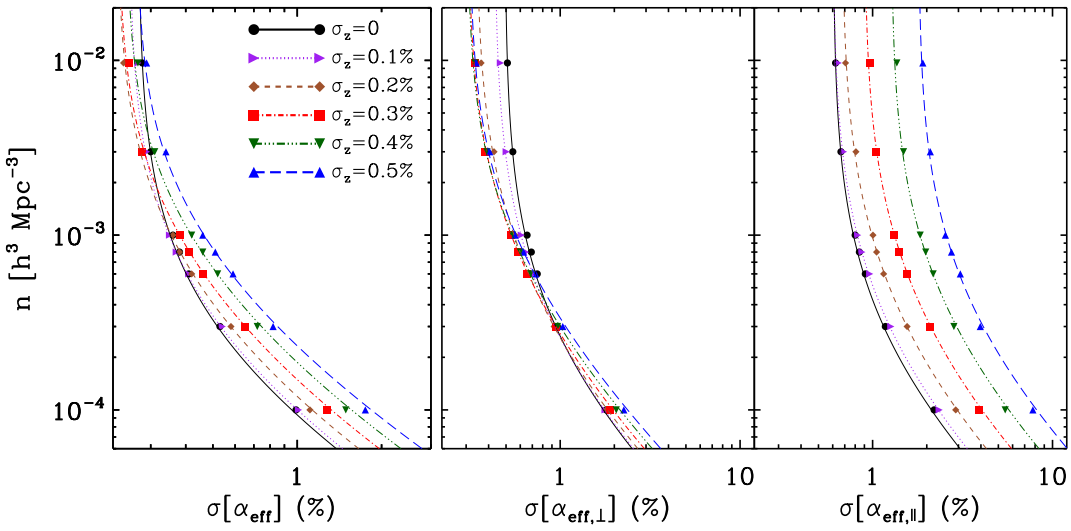


Figure 3.12: Precision with which $\alpha_{\rm eff}$ (left panel), $\alpha_{\rm eff,\perp}$ (middle panel), and $\alpha_{\rm eff,\parallel}$ (right panel) are extracted from samples with different number density and photo-z error. The symbols present the average results for the COLA ensemble and the lines shows the analytical predictions from Eq. 3.43 with $A = 0.215$. The black, magenta, brown, red, green, and blue colours indicate the results for samples with $\sigma_z = 0$, 0.1 %, 0.2 %, 0.3 %, 0.4 %, and 0.5 %. In Fig. 3.13 we present the constraints in Ω_m and ω derived for some of these samples.

σ_z. Moreover, a modest increase in the number density already delivers constraints comparable to those of the no error case. We also find that for $n = 3\times10^{-4}\,h^3\,{\rm Mpc}^{-3}$, the constraints on $\alpha_\perp$ for all the cases shown are almost identical.

The uncertainty in the parallel component of $\alpha_{\rm eff}$ grows with the photo-z error independently of the number density. This is the consequence of photo-z errors always reducing the weight of the parallel modes of the power spectrum when computing its angular average. For the parallel component, samples with photo-z errors need to considerably increase their number density to reach the same precision measuring $\alpha_\parallel$ as the achieved by samples with no errors. For instance, to produce the same constraints as samples with no errors and $n = 3\times10^{-4}\,h^3\,{\rm Mpc}^{-3}$, the number density have to be 3 and 5 times larger for $\sigma_z = 0.2\,\%$ and $\sigma_z = 0.3\,\%$, respectively. Moreover, for samples with $\sigma_z > 0.4\,\%$ and independently of their number density, it is impossible to achieve the same precision in the constraints as for samples with no errors and $n = 3\times10^{-4}\,h^3\,{\rm Mpc}^{-3}$.

In the three panels of Fig. 3.12 we show the predictions of our model for the uncertainty in $\alpha_{\rm eff}$ and its components (given by Eqs. 3.43 and 3.46) with coloured lines. To draw these lines we need the value of the normalization constant A, which we compute by fitting at different number densities $\sigma[\alpha_{\rm eff}]$ to our model. We obtain $A = 0.215$, and using this value, our model quantitatively captures the uncertainty in $\alpha_{\rm eff}$, where its precision is largely independent of the photo-z error and the number density of the sample.

All the above considerations should be taken into account for the optimal design

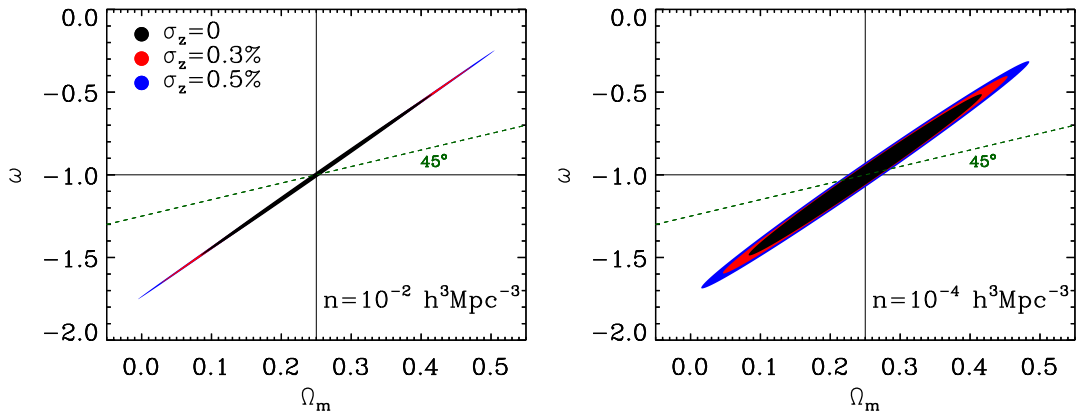


Figure 3.13: Constraints in Ω_m and ω derived from the average power spectrum monopole of the COLA ensemble with $n = 10^{-2}\,h^3\,\mathrm{Mpc}^{-3}$ (left panel) and $n = 10^{-4}\,h^3\,\mathrm{Mpc}^{-3}$ (right panel) and different photo-z errors. The black solid lines indicate the fiducial values of Ω_m and ω for the COLA simulations, and the green dashed lines show an angle of 45 degrees to guide the eye. These ellipses are characterised in Table 3.5. The uncertainty in Ω_m and ω after marginalising over the other parameters increases with the photo-z error, whereas the Figure of Merit (FoM) (inverse the ellipse area) is inversely proportional to the uncertainty in $\alpha_{\rm eff}$. Consequently, the sample with $\sigma_z = 0.3\,\%$ and $n = 10^{-2}\,h^3\,\mathrm{Mpc}^{-3}$ presents the greatest FoM.

of a survey or a target sample. For instance, the photo-z errors of a given galaxy sample might not only depend on the hardware employed, but also on the intrinsic galaxy properties (e.g. brighter objects having more accurate redshift estimates). In such case, the sample that delivers the strongest constraints in cosmological parameters is not necessarily that with the smallest photo-z error. We further explore the constraints in cosmological parameters for these samples in the following section.

3.5.4 Constraints in cosmological parameters

In the previous sections we showed that for $n > 3\times10^{-3}\,h^3\,\mathrm{Mpc}^{-3}$, samples with small photo-z errors ($\sigma_z < 0.5\,\%$) measure $\alpha_{\rm eff}$ with smaller uncertainty than samples with no errors. In this section we analyse how the precision with which $\alpha_{\rm eff}$ is measured translates into constraints in Ω_m and the equation of state of the dark energy.

In order to set constraints in the accelerated expansion of Universe, we have to choose a equation of state for the dark energy, $p = \omega\rho$, where p is the pressure, ρ is the energy density, and ω is a parameter that controls the equation of state. Here, we select this parameter to be redshift independent and with fiducial value $\omega = -1$.

In Fig. 3.13 we display the uncertainty in Ω_m and ω computed from the average power spectrum monopole of the COLA ensemble for different number densities and photo-z errors, as stated in the legend. The contours enclose the $1\,\sigma$ confidence region. Note that these constraints are derived by propagating the uncertainties in $\alpha_\parallel$ and $\alpha_\perp$, and assuming $\Omega_k = 0$ and ω constant. We characterize these elliptical contours together with the results for $n = 10^{-3}\,h^3\,\mathrm{Mpc}^{-3}$ in Table 3.5.

Table 3.5: Cosmological constraints from samples of DM particles with different number densities and photo-z errors.

$\sigma_z(\%)$	$a_{\rm ell}$	$b_{\rm ell}$	$\theta\,[^\circ]$	FoM
$n = 10^{-2}\,h^3\,{\rm Mpc}^{-3}$				
0	0.52	0.0054	71.27°	113.68
0.3	0.69	0.0036	71.27°	127.32
0.5	0.79	0.0036	71.27°	109.76
$n = 10^{-3}\,h^3\,{\rm Mpc}^{-3}$				
0	0.52	0.0070	71.26°	88.42
0.3	0.67	0.0057	71.27°	83.76
0.5	0.77	0.0060	71.27°	69.20
$n = 10^{-4}\,h^3\,{\rm Mpc}^{-3}$				
0	0.51	0.0192	71.21°	32.15
0.3	0.63	0.0203	71.22°	24.87
0.5	0.72	0.0241	71.22°	18.19

Notes. The parameters $a_{\rm ell}$ and $b_{\rm ell}$ present the value of the semi-major and semi-minor axes of the elliptical contours, where the contours enclose the $1\,\sigma$ confidence region for Ω_m and ω (see Fig. 3.13). The parameter $\phi = (-1)^N\,\theta + (\pi/2)\,N$, where N is a natural number, indicates the angle between these contours and the x-axis. The uncertainties in Ω_m and ω, after marginalising over the other parameter, are $\sigma[\Omega_m] \simeq a_{\rm ell}\,\cos\theta$ and $\sigma[\omega] \simeq a_{\rm ell}\,\sin\theta$.

The uncertainty in Ω_m and ω after marginalising over the other parameter increases with the photo-z error, whereas the Figure-of-Merit (FoM) of this combination of parameters is inversely proportional to the uncertainty in $\alpha_{\rm eff}$. Consequently, the FoM of samples with sub-percent photo-z errors may be greater than the one for samples without photo-z errors, as we showed in Fig. 3.12, e.g. for $n = 10^{-2}\,h^3\,{\rm Mpc}^{-3}$ and $\sigma_z = 0.3\,\%$ its value is FoM$\,=127$, whereas for the same number density and no photo-z errors it is FoM$\,=114$.

In §3.6 we estimate the FoM of this combination of parameters for the surveys that measure galaxy redshifts with sub-percent precision.

3.5.5 Effect of the PDF of photometric redshift errors

Along this work we have modelled photo-z errors as Gaussian distributions. However, this might not be necessarily a good approximation to reality under some circumstances. Therefore, to finalize this chapter, we explore the performance of our fitting procedure when considering probability distribution functions with different levels of skewness and excess kurtosis (see §3.2.3).

In Fig. 3.14 we present the average best-fit values of $\alpha_{\rm eff}$ extracted from samples of the COLA ensemble with $n = 10^{-2}h^3{\rm Mpc}^{-3}$ and different PDFs for which the

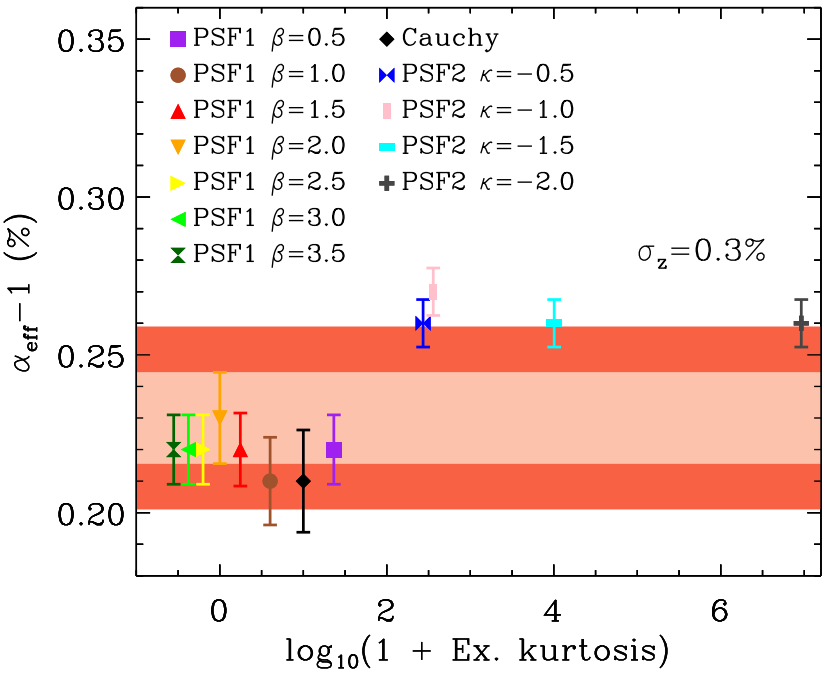


Figure 3.14: Average shift in $\alpha_{\rm eff}$ computed from the COLA ensemble for $n = 10^{-2}h^3{\rm Mpc}^{-3}$ and photo-z errors with different PDFs for which the difference between their 84th and 16th percentiles is $\sigma_z = 0.3\,\%$. The coloured areas enclose the $1\,\sigma$ and $2\,\sigma$ confidence regions for photo-z errors following a Gaussian distribution. The employed PDFs are introduced in §3.2.3. Even for photo-z errors drawn from PDFs with large excess kurtosis and skewness, the value of $\alpha_{\rm eff}$ after assuming a Gaussian PDF in the analysis is within $2\,\sigma$ from the value obtained for photo-z errors drawn from a Gaussian PDF.

difference between their 84th and 16th percentiles is $\sigma_z = 0.3\,\%$. Note that we plot the Cauchy distribution with an excess kurtosis of 9 because for this distribution the excess kurtosis is not defined. In addition, the shaded regions indicate the $1\,\sigma$ and $2\,\sigma$ confidence regions for a Gaussian PDF. The skewness for the family PDF1 is zero and for the family PDF2 monotonically increasing with κ, being 1.73 for $\kappa = 0.5$ and 416.9 for $\kappa = 2$. The bias in the stretch parameter is largely insensitive to the actual PDF shape at the statistical level of our simulated catalogues – all but one case is compatible with the Gaussian case at the $2\,\sigma$ level. For extreme PDFs, this could in principle introduce systematic errors in the estimation of $\alpha_{\rm eff}$. In practice and for the volume of future surveys[7], we expect that a reasonable estimate of the photo-z error PDF will produce unbiased results with respect to assuming a Gaussian PDF.

[7]Here, we are showing errors for the COLA ensemble, which encompasses $8100\,h^{-3}{\rm Gpc}^3$. At $z = 1$ and for the next generation of surveys, we expect approximately the same volume as for a single COLA simulation, $27\,h^{-3}{\rm Gpc}^3$.

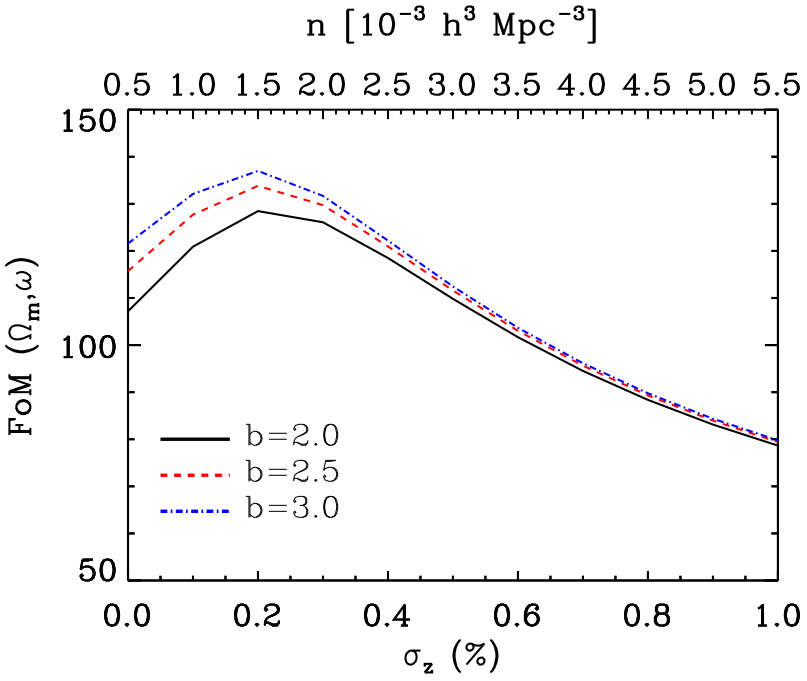


Figure 3.15: FoM with which the parameters Ω_m and ω can be measured for samples with different combinations of large-scale biases, photo-z errors, and number densities at $z = 1$. The colour of the line indicates the large-scale bias, as shown by the legend. The results are computed assuming that the number density linearly scales with σ_z and a volume of $V = 78.7\,\mathrm{Gpc}^3$. As we can see, there are samples with sub-percent photo-z errors that provide strictest constraints than spectroscopic samples.

3.6 Forecasts for future surveys with photo-z errors

In the previous sections we introduced a model for analytically estimating the uncertainty in $\alpha_{\rm eff}$ (see §3.4.3), which we contrasted with samples of DM particles and galaxies showing that it provides precise results independently of the large-scale bias, number density, and photo-z error of the sample. In addition, in §3.5.4 we estimated the precision that can be achieved in Ω_m and ω depending on the uncertainty in $\alpha_{\rm eff}$. Here, we will make forecasts for surveys that measure redshifts using noisy estimators.

In order to make forecasts for the FoM of Ω_m and ω for future surveys at $z = 1$, we need i) the effective volume covered by the survey, ii) the number density of galaxies as a function of the photo-z error, and iii) the average large-scale bias of the observed galaxies. We will assume that the survey observe a volume of $V = 78.7\,\mathrm{Gpc}^3$ at $z = 1$, i.e. the same volume as the MXXL and each COLA simulation, that the number of galaxies scales as $n = N(\sigma_z = 0)(\sigma_z\,10^3 + 1)$ (see table 8 of Benítez et al. 2014), where we assume that the number density of objects for a spectroscopic redshift at this redshift is $N(\sigma_z = 0) = 5 \times 10^{-4}\,h^3\,\mathrm{Mpc}^{-3}$. In addition, we will use different values for the average large-scale bias.

In Fig. 3.15 we show the results for $H_0 = 67.8\,\mathrm{km\,s^{-1}Mpc^{-1}}$, $\Omega_m = 0.308$, and

$\Omega_\Lambda = 0.692$ (Planck Collaboration et al. 2016a). As we can see, if the number density linearly scales with σ_z, the galaxy sample that provides the strictest constraints measuring Ω_m and ω is the one with $\sigma_z = 0.2\,\%$. This should not be surprising after looking at the left panel of Fig. 3.12, as the FoM is inversely proportional to the uncertainty in $\alpha_{\rm eff}$. In addition, we find that the higher the large-scale bias of the sample, the greatest the FoM, which is expected because a large bias reduces the contribution of the shot noise.

In conclusion, in order to design and fully exploit galaxy surveys that employ noisy estimators to compute photo-z errors, it is necessary to carefully select the properties of the target galaxy sample.

3.7 Conclusions

The next generation of cosmological surveys will dramatically increase the precision with which the expansion history of the Universe can be measured. Some of these future surveys will observe large areas of the sky with tens of narrow-bands and no preselection of the sources, providing a low-resolution spectra of every pixel of the sky. In addition, narrow-band surveys like J-PAS will reach a sub-percent redshift precision for hundreds of millions of galaxies, offering a new promising way of constraining cosmological parameters. Nevetheless, to fully exploit this new kind of data it is necessary to fully characterise the effect of photo-z errors on cosmological observables.

In this work we have presented a detailed study of the impact of photo-z errors on the clustering of galaxies, with an emphasis on the BAO signal. We have derived analytic expressions for how photo-z errors modify the power spectrum multipoles, their variances, and the amplitude of the BAO, which we have contrasted with results from hundreds of N-body simulations.

Our main findings can be summarised as follows:

- We showed analytically and with simulations that galaxy samples with large number densities and photo-z errors have a higher SNR in the monopole and quadrupole than samples with no photo-z errors on scales of the order of $k\,\sigma_{\rm eff} \simeq 1$ and $k\,\sigma_{\rm eff} \simeq 1.5$, respectively. This is important because for samples with sub-percent photo-z errors these scales correspond to the scales where the BAO are located.
- In the left-panel of Fig. 3.7 we displayed an analytic expression for the suppression of the BAO in the monopole as a function of the photo-z error and the scale. We found that the sharpness of the BAO grows with the photo-z error, converging to the BAO smearing for modes perpendicular to the light-of-sight, which suffer the smallest suppression. This is because photo-z errors reduce the weight of power spectrum modes parallel to the line-of-sight when computing the angular average, where these modes are more suppressed due to RSD. We confirmed these results with measuring the BAO from hundreds of N-body simulations.
- We derived how the cosmological information encoded on the BAO depends on the number density, photo-z error, and large-scale bias of the galaxy sample. We showed that small-scale RSD and/or photo-z errors induce a scale-dependence

on the cosmological information encoded in the BAO feature, where large scales set stronger constraints on the Hubble parameter and small scales on the angular diameter distance.

- Based on these findings, we built a model for extracting the BAO information from the monopole. We then applied this model to simulated galaxy catalogues with different levels of shot noise, large-scale bias, and photo-z errors. In Fig. 3.10 we showed that photo-z do not significantly shift the position of the BAO with respect to the case with no photo-z errors. Moreover, we found that the precision with which the BAO scale can be extracted depends on a balance of three effects. The first is that photo-z errors increase the SNR in the monopole on scales where $k\,\sigma_{\rm eff} \simeq 1$, the second is that the suppression of the BAO decreases with σ_z, and the third that the scale at which the shot noise dominates the monopole grows with σ_z. As a consequence, for large number densities we found stricter constraints on $\alpha_{\rm eff}$ from simulated catalogues with sub-percent photo-z errors than from simulated catalogues with no errors.

- In §3.5.4 we analysed how the uncertainty in $\alpha_{\rm eff}$ is translated into the precision measuring Ω_m and ω, and we computed the value of the FoM of these parameters as a function of the number density and the photo-z error of the sample. We found that for large number densities, the precision measuring Ω_m and ω is greater for samples with sub-percent photo-z errors than for samples with no errors.

In summary, we have proved that it is crucial a profound understanding of the effect of photo-z errors on the galaxy clustering to extract correctly the BAO scale and to determine the cosmological information encoded in this scale. Moreover, we have showed that it is very important to characterise the number density, photo-z error, and large-scale bias of the galaxy sample employed to measure the BAO scale, where smaller photo-z errors do not always increase the precision measuring $\alpha_{\rm eff}$. To explore the constraints on cosmological parameters as a function of the large-scale bias, number density, and photo-z, in §3.4.3 we introduced a simple model that provides $\sigma[\alpha_{\rm eff}]$ for galaxy samples with different properties and distinct cosmologies.

Appendix A: Expressions for the effect of photo-z errors on $P(k)$

The expressions needed to compute the variance of the power spectrum monopole and quadrupole for a Gaussian $\Pr(\delta z)$ are the following:

$$\begin{aligned}\left\langle \mathcal{F}^4\right\rangle =&\frac{\sqrt{\pi}}{2}\frac{\mathrm{Erf}(x)}{x}\left(1+\frac{2\beta}{x^2}+\frac{9\beta^2}{2x^4}+\frac{15\beta^3}{2x^6}+\frac{105\beta^4}{16x^8}\right)\\&-\frac{2\beta\exp(-x^2)}{x^2}\left(1+\frac{9\beta}{4x^2}\mathcal{H}_1(x)+\frac{15\beta^2}{4x^4}\mathcal{H}_2(x)+\frac{105\beta^3}{32x^6}\mathcal{H}_3(x)\right),\end{aligned} \tag{3.50}$$

$$\begin{aligned}\left\langle \mu^2\mathcal{F}^4\right\rangle =&\frac{\sqrt{\pi}}{4}\frac{\mathrm{Erf}(x)}{x^3}\left(1+\frac{6\beta}{x^2}+\frac{45\beta^2}{2x^4}+\frac{105\beta^3}{2x^6}+\frac{1890\beta^4}{x^8}\right)\\&-\frac{\exp(-x^2)}{2x^2}\left(1+\frac{6\beta}{x^2}\mathcal{H}_1(x)+\frac{45\beta^2}{2x^4}\mathcal{H}_2(x)\right.\\&\left.+\frac{105\beta^3}{2x^6}\mathcal{H}_3(x)+\frac{1890\beta^4}{x^8}\mathcal{H}_4(x)\right),\end{aligned} \tag{3.51}$$

$$\begin{aligned}\left\langle \mu^4\mathcal{F}^4\right\rangle =&\frac{3\sqrt{\pi}}{8}\frac{\mathrm{Erf}(x)}{x^5}\left(1+\frac{10\beta}{x^2}+\frac{105\beta^2}{2x^4}+\frac{315\beta^3}{2x^6}\right.\\&\left.+\frac{3465\beta^4}{16x^8}\right)-\frac{3\exp(-x^2)}{4x^4}\left(\mathcal{H}_1(x)+\frac{10\beta}{x^2}\mathcal{H}_2(x)\right.\\&\left.+\frac{105\beta^2}{2x^4}\mathcal{H}_3(x)+\frac{315\beta^3}{2x^6}\mathcal{H}_4(x)+\frac{3465\beta^4}{16x^8}\mathcal{H}_5(x)\right),\end{aligned} \tag{3.52}$$

where $x=\sqrt{2}\,k\,\sigma_{\rm eff}$. We only employ the previous expressions together with the ones provided in the main body of the text when $x>3$ and $u>3$, as they diverge as $x\to 0$ and $u\to 0$. In order to derive precise expressions for $x\leq 3$ and $u\leq 3$, we expand the exponential functions into power series, obtaining:

$$\left\langle \mathcal{F}^2\right\rangle =\sum_{j=0}^{\infty}(-1)^j\frac{x^{2j}}{j!}\left(\frac{1}{2j+1}+\frac{2\beta}{2j+3}+\frac{\beta^2}{2j+5}\right), \tag{3.53}$$

$$\left\langle \mu^2\mathcal{F}^2\right\rangle =\sum_{j=0}^{\infty}(-1)^j\frac{x^{2j}}{j!}\left(\frac{1}{2j+3}+\frac{2\beta}{2j+5}+\frac{\beta^2}{2j+7}\right), \tag{3.54}$$

$$\left\langle \mathcal{F}^4\right\rangle =\sum_{j=0}^{\infty}(-2)^j\frac{x^{2j}}{j!}\left(\frac{1}{2j+1}+\frac{4\beta}{2j+3}+\frac{6\beta^2}{2j+5}+\frac{4\beta^3}{2j+7}+\frac{\beta^4}{2j+9}\right), \tag{3.55}$$

$$\left\langle \mu^2\mathcal{F}^4\right\rangle =\sum_{j=0}^{\infty}(-2)^j\frac{x^{2j}}{j!}\left(\frac{1}{2j+3}+\frac{4\beta}{2j+5}+\frac{6\beta^2}{2j+7}+\frac{4\beta^3}{2j+9}+\frac{\beta^4}{2j+11}\right), \tag{3.56}$$

$$\left\langle \mu^4\mathcal{F}^4\right\rangle =\sum_{j=0}^{\infty}(-2)^j\frac{x^{2j}}{j!}\left(\frac{1}{2j+5}+\frac{4\beta}{2j+7}+\frac{6\beta^2}{2j+9}+\frac{4\beta^3}{2j+11}+\frac{\beta^4}{2j+13}\right), \tag{3.57}$$

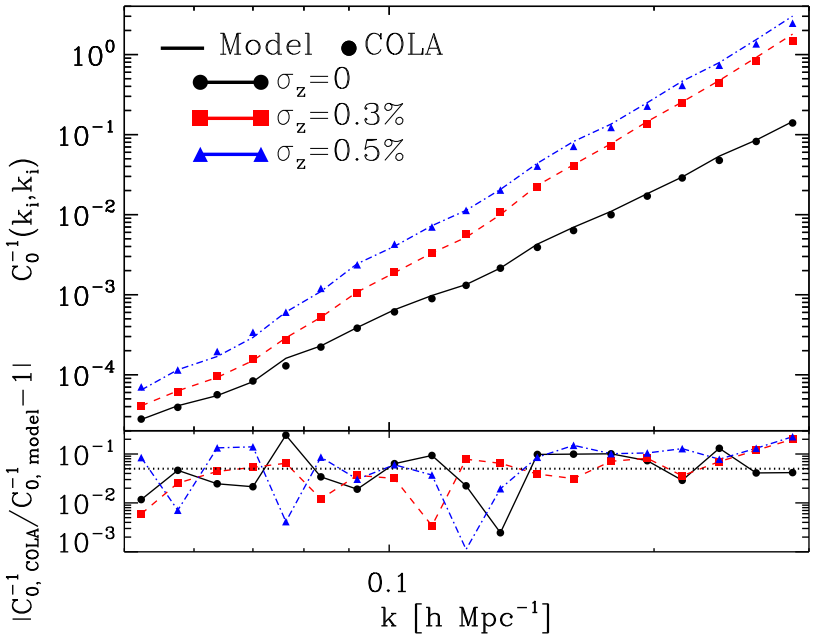


Figure 3.16: Diagonal terms of the precision matrix of the redshift-space power spectrum monopole for different photo-z errors. The lines indicate analytic results obtained by inverting Eq. 3.20 and the points show precision matrices estimated from the COLA ensemble after correcting for a prefactor introduced by Hartlap et al. (2007). The bottom panel show that the relative difference of the COLA and analytic precision matrices, where the average precision is $\simeq 6\,\%$.

$$\left\langle G(k,\mu)\mathcal{F}^2\right\rangle = e^{-\frac{1}{2}(k\,\sigma_\perp)^2}\sum_{j=0}^{\infty}(-1)^j\frac{u^{2j}}{j!}\left(\frac{1}{2j+1}+\frac{2\beta}{2j+3}+\frac{\beta^2}{2j+5}\right), \tag{3.58}$$

where summing from $j=0$ to $j=90$ is more than enough to achieve 6 significant digits of precision.

Appendix B: Effect of off-diagonal terms on the covariance matrix

In §3.3 we analytically derived the diagonal terms of the power spectrum monopole covariance matrix, showing that photo-z do not increase the value of the off-diagonal terms. In addition, we showed that the off-diagonal terms of the covariance matrix are small with respect to the diagonal terms on the scales where the BAO are located. Here, we explore the importance of these terms when extracting the BAO scale.

In Fig. 3.16 we show the diagonal terms of the precision matrices estimated from the COLA ensemble and analytic precision matrices computed by inverting Eq. 3.20. The first are indicated by points and the second by lines, where their colour denote the photo-z error, as indicated by the legend. In the bottom panel we present the

relative difference between the analytic precision matrices and the ones computed using the COLA ensemble. We find that the average precision of the model is $\simeq 6\,\%$. Consequently, this model is able to reproduce the diagonal terms of the precision matrices.

To check whether the off-diagonal terms of the precision matrices are important when extracting the BAO scale, we repeat the same analysis as in §3.5 using analytic precision matrices. We find that the relative difference between the new and old values of $\alpha_{\rm eff}$ is to within $4\,\%$ and thus, the effect of off-diagonal terms at $z = 1$ is sub-dominant. All of this motivates the use of analytic precision matrices, since they can be instantly computed for different combinations of photo-z errors, large-scale biases, number densities, and cosmological parameters.

4 Identification and redshift estimation of high-z AGN

"Now there's a look in your eyes, like black holes in the sky".

— Pink Floyd, *Shine On You Crazy Diamond*

4.1 Introduction

AGN are among the brightest objects in the Universe. They are powered by the accretion of matter onto a SMBH: as the gas approaches the SMBH, its temperature rises and starts to emit radiation across the entire electromagnetic spectrum. Nevertheless, AGN not only show a continuum emission from the gas in the accretion disk, they also exhibit multiple emission lines from the X-ray to the infrared spectral range. In turn, the emission lines may be broad or narrow, depending on the orientation of the AGN with respect to the observer and the obscuring material (AGN unification scheme, Antonucci 1993; Urry & Padovani 1995). They are also employed to classify AGN from observations, sources with broad and narrow lines are named type-I AGN and objects with just narrow lines are called type-II AGN.

For their many applications in different fields of astrophysics, from high-energy physics to cosmology, a complete census of AGN is fundamental. They may be employed to constrain galaxy evolution models (e.g., Heckman & Best 2014), as there are hints of correlations between SMBH and galaxy properties (e.g., Kormendy & Richstone 1995; Gebhardt et al. 2000), whereas a causal origin of this correlation is not universally accepted (e.g., Peng 2007; Jahnke & Macciò 2011). Moreover, thanks to their large luminosity, the optically brightest type-I AGN, commonly referred to as quasars, allow us to trace the matter distribution since early times (currently, the most distant spectroscopically-confirmed quasar is at $z = 7.1$, see Mortlock et al. 2011). They can also be used to measure cosmology: Busca et al. (2013) successfully detected BAO in the Ly α forest and future galaxy surveys will employ their distribution to measure BAO (e.g., eBOSS is expected to reach a 1.6 % precision measuring spherically averaged BAO with them, see Dawson et al.

2016; Zhao et al. 2016). Finally, they have even been proposed as standard candles (Wang et al. 2014; Watson et al. 2011; Risaliti & Lusso 2016).

There are many techniques for detecting AGN, such as traditional colour-colour selections (Matthews & Sandage 1963); intrinsic variability in the optical (Schmidt et al. 2010); and the combination of multi-wavelength data, like radio (e.g., White et al. 2000), X-ray (e.g., Barger et al. 2003; Brusa et al. 2003), and infrared (e.g., Lacy et al. 2004). The strengths and weaknesses of these methods are different. For instance, X-ray surveys are very time consuming and they produce pure and complete samples of AGN, whereas optical images are less time expensive and capable of detecting many AGN. However, optical images are biased towards unobscured type-I AGN.

The emergence of medium- and narrow-band photometric surveys, such as COMBO-17 (Wolf et al. 2004, 2008), COSMOS (Ilbert et al. 2009), the ALHAMBRA survey (Moles et al. 2008), SHARDS (Pérez-González & Cava 2013), PAUS (Martí et al. 2014), and the upcoming J-PAS (Benítez et al. 2014), open the possibility of exploring new methods for detecting AGN. They produce multi-band photometric data that combines the strengths photometric and spectroscopic surveys, resulting in a low-resolution spectra of every pixel of the sky. The aim of this work is precisely to produce a new pipeline to identify AGN and to compute their redshifts. In order to do this, we take advantage of the data from medium- and narrow-band surveys to identify strong spectral features typical of active galaxies.

We test our new algorithm, dubbed as ELDAR, by applying it to the data from the ALHAMBRA survey (Moles et al. 2008; Molino et al. 2014). This survey is an optimal test-case for ELDAR because it observed $\simeq 4\,\mathrm{deg}^2$ using 20 contiguous medium-band filters (FWHM $\simeq 300\,$Å) in the optical range and 3 broad-band filters (J, H, and Ks) in the infrared. We extract two catalogues of type-I AGN using two different ELDAR configurations, the first maximising completeness and the second minimising contamination. Then, we analyse the main properties of these catalogues and we estimate their completeness, redshift precision, and galaxy contamination by applying the same ELDAR configurations to samples of spectroscopically-known type-I AGN and galaxies within the ALHAMBRA fields.

This chapter is structured as follows. In §4.2 we introduce ELDAR and in §4.3 we tune our method to detect type-I AGN in ALHAMBRA. In §4.4 we extract two catalogues of type-I AGN employing ELDAR, and we characterize their properties using samples of spectroscopically-known type-I AGN and galaxies within the ALHAMBRA fields. In §4.5 we discuss the potential of our methodology for surveys with narrower bands and in §4.6 we summarise our conclusions.

Throughout this chapter we simply refer to all classes of active galaxies as AGN, to active galaxies with broad emission lines as type-I AGN, and to active galaxies with just narrow emission lines as type-II AGN. The optical and near-IR magnitudes are in the AB system, we always employ the spectral flux density per unit wavelength, and we assume a ΛCDM cosmology with $H_0 = 67.8\,\mathrm{km\ s^{-1}Mpc^{-1}}$, $\Omega_\Lambda = 0.692$, and $\Omega_\mathrm{M} = 0.308$ (Planck Collaboration et al. 2016a).

4.2 ELDAR algorithm

ELDAR consists of two main steps: i) *template fitting*, that aims at pre-selecting AGN candidates and obtaining a Redshift Probability Distribution Function (PDZ) for each one of them, and ii) *spectro-photometric confirmation*, whose objective is to securely confirm the previous candidates by detecting AGN emission lines and to refine the photo-z estimation.

In what follows, we describe in more detail the two steps of our methodology.

4.2.1 Template fitting step

The objective of this first step is to pre-select AGN candidates, and to obtain a PDF for each of them. While any template-fitting code or machine learning algorithm may be used for this pre-selection phase, in this work we adopted PHotometric Analysis for Redshift Estimate (LePHARE) (Arnouts et al. 1999). LePHARE is a template-fitting code extensively used to compute photo-zs for galaxies and AGN (e.g., Ilbert et al. 2009; Salvato et al. 2009, 2011; Fotopoulou et al. 2012; Matute et al. 2012). Here we provide a general discussion on how to correctly configure LePHARE for detecting AGN. This is because the templates and parameters of the code have to be carefully chosen and optimised depending on the characteristics of the survey to be analysed (in §4.3.3, we provide the specific configuration of LePHARE for the case of the ALHAMBRA survey).

- *Template selection.* LePHARE classifies each source and computes its redshift depending on the Spectral Energy Distribution (SED) of the template that produces the best-fit to its photometric data, where a template is a theoretical or empirical curve that describes the flux of astronomical objects as a function of λ. The library of templates to be used LePHARE has to be meticulously chosen, especially when working with AGN (Hsu et al. 2014). While the library should be comprehensive enough to include the broad variety of SEDs of the types of sources that are sought, the number of templates should not be too large so as to avoid degeneracies.

  The templates are divided into two categories in LePHARE: stellar and extragalactic. The first includes the SEDs of stars, while the second presents the SEDs of extragalactic objects at rest-frame, which are shifted in redshift during the fitting procedure. To build our stellar library we include 254 stellar templates from the publicly available distribution of LePHARE. They are divided into 131 templates of normal stellar spectral types and luminosity classes at solar abundance, metal-poor F-K dwarfs, and G-K giants (Pickles 1998); 4 templates of white dwarfs (Bohlin et al. 1995); 100 templates of low mass stars (Chabrier et al. 2000); and 19 sub-dwarfs (Bixler et al. 1991). We include all of them to cover as many stellar types as possible.

  For the extragalactic library, we only include templates of active galaxies, as these are the sources we are targeting. With this approach, we ensure that no AGN are wrongly classified as a 'normal' galaxies, i.e. galaxies whose SED is not dominated by the nuclear activity, while all normal galaxies will be discarded by the spectro-photometric confirmation step (see §4.2.2). The AGN templates

to be included are survey specific as the AGN types that can be unambiguously detected depend on the characteristics of the survey to be analysed, e.g. its depth, area, and the width of its photometric bands. In particular, the width of the bands determines the approximate minimum Equivalent Width (EW) of the emission lines that can be detected by ELDAR (see §4.3.3). As the EW of AGN emission lines depends on the type of active galaxy, we should only include templates of AGN with emission lines strong enough to be detected by ELDAR.

- *Redshift range and precision.* The extragalactic templates included in the LePHARE library are located at rest-frame. During the fitting procedure, LePHARE creates a grid of templates within a redshift range defined by the user. As Benítez et al. (2009b) observed, the size of the redshift step should depend on the number of filters available and how close they are to each other. As for the maximum redshift, we set it to the redshift above which no strong spectral features are presented to within the medium- or narrow-band wavelength coverage.

  Effectively, the PDZ generated by LePHARE is defined as:

$$\mathrm{PDZ}(z) = \frac{G(z)}{G(z_{\mathrm{best}})}, \tag{4.1}$$

  where $G(z) = \exp[-\chi^2_{\mathrm{min}}(z)/2]$, $\chi^2_{\mathrm{min}}(z)$ is the χ^2 resulting from the template that best fits the data at redshift z, and z_{best} is the redshift at which the data is best fitted. With this definition, the PDZ is not properly a probability density function and to generate one for each object, the PDZ should be normalized by its integral.

- *Dust attenuation.* The extinction law of AGN varies as a function of redshift (e.g., Gallerani et al. 2010), reflecting different mechanisms for dust production and/or destruction. A correct modelling of the effect of dust is required because the dust absorbs UV and optical light, which then emits in the infrared modifying the SED. We employ the Milky Way (Allen 1976), Small Magellanic Cloud (Prevot et al. 1984), Large Magellanic Cloud (Fitzpatrick 1986), and starburst (Calzetti et al. 2000) extinction laws, which are shown in fig. 7 of Bolzonella et al. (2000).

  The dust attenuation (A_V) depends on the orientation of the AGN and it is defined as

$$A_V = R_V \times E(B-V), \tag{4.2}$$

  where $E(B-V)$ is the colour excess and R_V is a coefficient that depends on the extinction law. We introduce colour excesses from 0 to 0.10 in steps of 0.02, from 0.10 to 0.30 in steps of 0.04, and from 0.30 to 1.00 in steps of 0.10. We include colour excesses as high as 1 to account for very extinguished AGN. We set finer steps for low colour excesses because some AGN templates are empirical, and thus they already include some extinction.

- *Luminosity prior.* Setting luminosity priors is important to avoid unrealistic solutions (Salvato et al. 2009) and they should be chosen depending on the type of objects that we want to target. Quasars, for example, are traditionally defined as objects with $M_B \leq -23$ and setting $M_B = -23$ as upper limit ensures that LePHARE rejects low redshift (low-z) solutions.

4.2.2 Spectro-photometric confirmation step

Objects with strong emission lines, such as type-I AGN, are particularly suited to be detected in surveys with medium- and/or narrow-band filters. This is because emission lines with large EW completely dominate the bands in which they fall, and these bands appear as clear 'peaks' in the photometric SED. The relative height of these peaks depends on i) the EW of the line, ii) the width of the band where the emission line falls, and iii) the shape of the continuum emission. If we assume the that flux per unit of wavelength of the AGN continuum emission is flat in the bands adjacent to the band where the line falls, ELDAR is able to detect lines with EW greater than

$$\mathrm{EW}_{\mathrm{min}} = \frac{B_{\mathrm{FWHM}}}{(1+z)B_{\mathrm{SNR}}}\sigma_{\mathrm{line}}, \tag{4.3}$$

where B_{FWHM} is the full width half maximum of the band in which the line falls, B_{SNR} is the SNR in this band, z is the redshift of the source, and σ_{line} is a parameter that denotes the confidence with which we want to confirm lines, e.g. $\sigma_{\mathrm{line}} = 1$ means a $1\,\sigma$ detection. Therefore, the detection of emission lines depends on the intrinsic properties of each source, e.g. its redshift and the strength of its emission lines, and on the characteristics of the survey to be analysed, such as the width of its bands and its depth.

We note; however, that Eq. 4.3 just sets an ideal value of $\mathrm{EW}_{\mathrm{min}}$ because the assumption of a flat continuum is generally not correct for AGN, especially at $z < 2.5$ where the slope of the AGN continuum is usually very steep and blue. Moreover, for emission lines broader that the survey bands and falling in between two bands, the flux is dispersed, and thus the value of $\mathrm{EW}_{\mathrm{min}}$ decreases too.

With these caveats in mind, the objective of this second step of ELDAR is precisely to search for typical AGN emission lines in the SED of the sources that we want to classify. Therefore, we significantly improve on the ability of template fitting codes in detecting emission line objects, as they do not include special weights in the bands where emission lines fall and, as the number of bands dominated by the continuum is always greater than the number of bands dominated by emission lines in medium- and narrow-band surveys, they are not specifically designed for detecting these objects.

The detection of AGN emission lines allows not only the confirmation of objects as active galaxies but also the rejection of wrongly-classified sources, such as stars and galaxies. Moreover, it provides a method to discriminate between different redshift solutions given by the PDZ. Operationally, the confirmation step works as follows:

1. We start by selecting, for each source, the redshifts at which the SED is best fitted by an extragalactic template ($\chi^2_{\mathrm{AGN}} < \chi^2_{\mathrm{star}}$) and the value of the PDZ

is greater than 0.5. We set a lower limit in the PDZ in order to include the information provided by LePHARE from the fitting of the SED. We check the dependence of the results on different PDZ lower limits in Appendix B. For each of these possible redshift solutions, $z_{\rm guess}$, we perform the steps that follow.

2. According to each $z_{\rm guess}$, we calculate which AGN emission lines with EW greater than $\mathrm{EW}_{\rm min}$ are expected within the wavelength coverage of the survey, and in which band they should fall. We then confirm the detection of a line if:

$$F_{\rm cen} > \begin{cases} F_{\rm blue} + \sigma_{\rm line}\, S_{\rm cen}, \\ F_{\rm red} + \sigma_{\rm line}\, S_{\rm cen}, \\ F_{\rm blue} + \sigma_{\rm line}\, S_{\rm blue}, \\ F_{\rm red} + \sigma_{\rm line}\, S_{\rm red}, \end{cases} \tag{4.4}$$

where $F_{\rm cen}$ is the flux in the band where the line should fall, $F_{\rm blue}$ ($F_{\rm red}$) is the flux in the band bluewards (redwards) to the band where the line should fall, and $S_{\rm cen}$, $S_{\rm blue}$, and $S_{\rm red}$ are their errors. By construction, we are unable to confirm lines that fall either in the first or in the last band of the survey filter system.

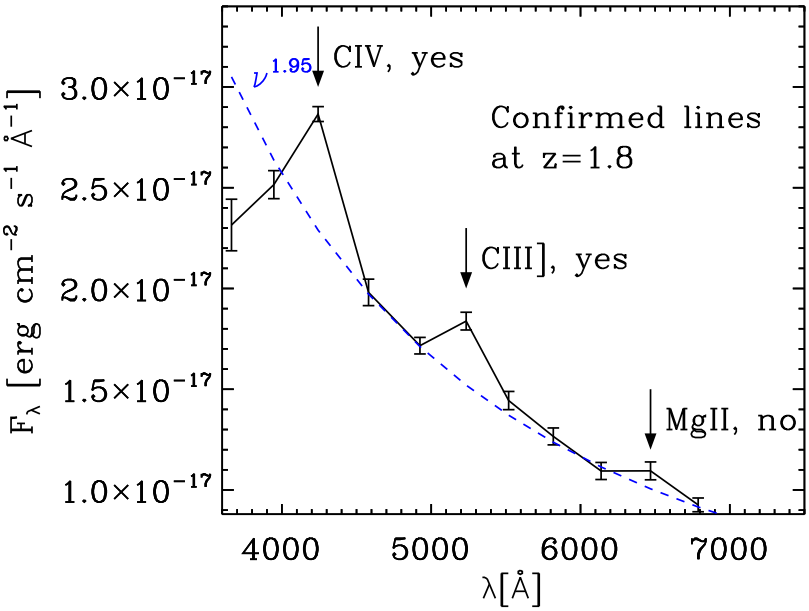


Figure 4.1: Multi-band ALHAMBRA photometry of a spectroscopically-known type-I AGN at $z = 1.8$. At this redshift, the lines C IV, C III], and Mg II fall within the ALHAMBRA medium-band wavelength range. ELDAR confirms C IV and C III] with more than $1\,\sigma$ confidence in the 3rd and 6th band, respectively. On the other hand, Mg II is not confirmed because the flux in the 9th band, where this line should fall according to $z_{\rm spec}$, does not fulfil all the requirements of Eq. 4.4. The blue dashed line shows a power law to guide the eye on the AGN continuum.

In Fig. 4.1 we show a spectroscopically-known type-I AGN at $z_{\rm spec} = 1.8$ observed by the ALHAMBRA survey (we will present it in §4.3.1). We show

arrows pointing to the bands where C IV, C III], and Mg II should fall according to $z_{\rm spec}$. The blue dashed line indicates a power law that guides the eye on the continuum emission and it allows us to easily see the flux excess in the bands where the emission lines fall. In this example, C IV and C III] are detected by ELDAR while Mg II is not confirmed because it does not fulfil all the requirements of Eq. 4.4.

There are some redshift intervals for which two different emission lines may fall in consecutive bands, and thus the line detection is not secure. However, the typical separation between the strongest AGN emission lines (EW > 8) with rest-frame central wavelength $\lambda_{\rm c} < 4\,000$ Å is large enough for these lines to never fall in consecutive bands in surveys with bands narrower than FWHM ~ 400 Å. In addition, if lines with different EW fall in consecutive bands, the line with the largest EW can still be confirmed.

In surveys with no contiguous bands another complication might arise at redshifts in which AGN emission lines fall between two bands, as the flux of the line gets dispersed. However, in most cases the greatest part of the line falls in one band and just its tail in others. In this case, the line is detected in the band where the greatest part of the line falls. We further explore this issue in §4.5.

To account for redshift errors and physical processes, such as line shifts and anisotropic profiles (see Vanden Berk et al. 2001), that may displace emission lines from the band where they should fall, we search for them not only in the band where they should fall according to $z_{\rm guess}$, but also in the adjacent bands.

3. We confirm as AGN the sources for which we detect at least $\mathcal{N}$ emission lines, where $\mathcal{N}$ should be chosen depending on the number of lines that the filter system allows to detect, as well as on a compromise between the purity and completeness to be achieved. Obviously, the larger the number of lines detected, the higher the purity of the resulting catalogue (see §4.3.2 for a discussion about potential contaminants for the ALHAMBRA survey).

4. Once an AGN is confirmed, we check at which $z_{\rm guess}$ the largest number of lines is detected, rejecting the other values. If we end up with a single $z_{\rm guess}$, we accept it as the final photo-z solution, $z_{\rm phot}$. Otherwise, we group contiguous $z_{\rm guess}$ into intervals, and we look for the interval with the greatest average PDZ. Finally, we compute the final redshift solution as

$$z_{\rm phot} = \frac{\sum_i^n z_{{\rm guess},i}\,{\rm PDZ}(z_{{\rm guess},i})}{\sum_i^n {\rm PDZ}(z_{{\rm guess},i})}, \tag{4.5}$$

where the summation goes through the n values of the $z_{\rm guess}$ in the selected interval.

In Fig. 4.2 we show an illustrative example of this procedure. We start by selecting $z_{\rm guess}$, i.e. the redshifts at which the SED of the object is best fitted by an AGN template and the value of the PDZ is greater than 0.5, which are the red, green, and blue points. Then, we pick the $z_{\rm guess}$ for which the largest number of lines is detected (in this example, the red and blue dots). After that, we group the red points

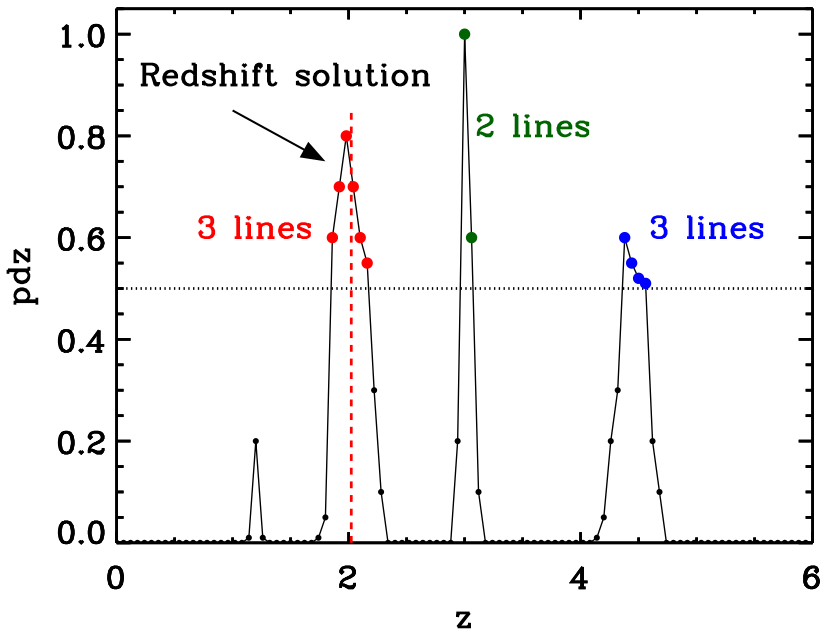


Figure 4.2: Mock example of a PDZ and information about the number of AGN emission lines detected by ELDAR. The black small dots indicate redshift solutions with PDZ < 0.5, the green dots the solutions with PDZ > 0.5 and for which ELDAR detects 2 AGN emission lines, and the red and blue dots the solutions with PDZ > 0.5 and for which ELDAR detects 3 AGN emission lines. The red dashed line shows the final redshift solution for the source, $z_{\rm phot}$. See the text for further information about how $z_{\rm phot}$ is computed.

into one redshift interval and the blue ones into another. We then reject the blue-points interval because the mean PDZ of the red-points interval is greater. Finally, we compute $z_{\rm phot}$ with the red-points interval using Eq. 4.5.

The above steps define the backbone of the spectro-photometric confirmation. Additional criteria can be added to refine the procedure. For instance, as the Ly α line is the strongest AGN emission line in the UV, in the present work we require i) that the Ly α line has to be detected for sources with redshift solutions for which this line should fall within the wavelength coverage of the survey, and ii) that the flux in the band where it falls has to be at least 75 % of the maximum flux in any of the other bands. Even the Ly α line is the strongest in the UV, we set a 75 % limit to account for the possibility of the line falling in between two bands and other emission lines surpassing its flux. This condition aims at rejecting cold stars for which their continuum emission may be confused with the Lyman-break of high-z AGN. We explore the dependence of the results on this criterion in Appendix B.

4.3 Applying ELDAR to ALHAMBRA data

In the previous section we introduced ELDAR, our procedure for detecting AGN. Here, we apply ELDAR to the ALHAMBRA survey, with the objectives of i) testing the effectiveness of our method, and ii) extract photometric samples of type-I AGN from ALHAMBRA. We start by introducing the ALHAMBRA survey, and then we discuss some effects that may reduce the quality of ELDAR's results. After that, we show how we have optimized ELDAR for analysing the ALHAMBRA data and we finally summarise the main aspects of our methodology.

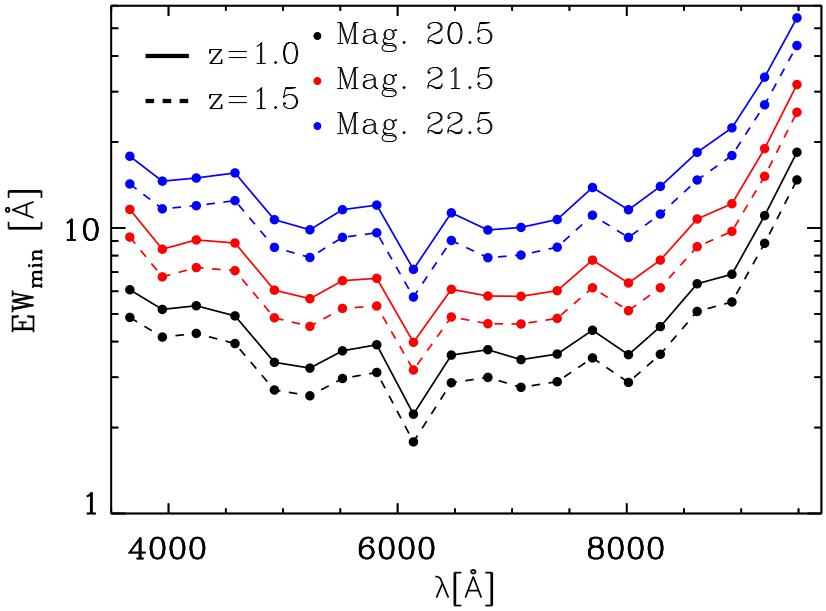


Figure 4.3: Minimum EW of emission lines that can be detected in each ALHAMBRA medium-band for $\sigma_{\rm line} = 1$, as a function of the magnitude of the band and the redshift of the source.

4.3.1 The ALHAMBRA survey

ALHAMBRA[1] is a medium-band photometric survey that observed $\simeq 4\,\mathrm{deg}^2$ distributed over 8 non-overlapping fields. These fields were selected to be in common with other surveys, such as Deep Extragalactic Evolutionary Probe 2 (DEEP2), SDSS, COSMOS, Hubble Deep Field North (HDF-N), Deep Groth Strip Survey (GROTH), and European Large Area ISO Survey (ELAIS). The ALHAMBRA filter system consists of 20 contiguous medium-band filters of width $\simeq 300\,$Å, which cover the optical range from 3500 to 9700 Å, and the 3 broad-band infrared filters J, H, and $K_{\rm s}$. The magnitude limit ($5\,\sigma$, 3") is $\simeq 23.7$ for the blue optical filters, $\simeq 22.2$ for the red optical filters, and $\simeq 22$ for the infrared filters (Aparicio Villegas et al. 2010). Due to the

[1] http://www.alhambrasurvey.com

Table 4.1: Emission lines employed to confirm type-I AGN in ALHAMBRA. At least 2 and 3 emission lines must be detected to validate objects using the ELDAR's 2- and 3-lines mode, respectively (see §4.3.3).

Line	λ_c(Å)	$\langle$EW(Å)$\rangle$
O VI+Ly β	1030	15.6±0.3
Ly α	1216	91.8±0.7
Si IV+O IV]	1397	8.13±0.09
C IV	1549	23.8±0.1
C III]	1909	21.2±0.1
Mg II	2799	32.3±0.1

Notes. The values of the central wavelengths and EWs are taken from Telfer et al. (2002) (for lines with $\lambda_c < 1300$Å) and from Vanden Berk et al. (2001) (for lines with $\lambda_c > 1300$Å).

width of its filters, and the contiguous coverage from the near UV to the near-infrared, the ALHAMBRA survey is an optimal test-case for ELDAR.

The last public data release of ALHAMBRA is introduced in Molino et al. (2014) (M14). It covered an area of $\simeq 3\,\mathrm{deg}^2$ over 7 fields, detecting 438 356 sources brighter than 24.5 mag in the synthetic detection band, F814W. This band was generated by combining the 9 reddest ALHAMBRA bands to mimic the *Hubble Space Telescope* (HST) - Advanced Camera for Surveys (ACS) F814W band.

The ALHAMBRA filter system produces precise redshift estimates for blue and red galaxies, as shown by M14. Specifically, M14 found a redshift precision of $\sigma_z \simeq 1\,\%$ for spectroscopically-known galaxies with F814W < 22.5 within the ALHAMBRA fields. Moreover, in a first attempt to characterise the ability of ALHAMBRA to produce precise photo-zs for type-I AGN, Matute et al. (2012) applied LePHARE to a sample of 170 spectroscopically-known type-I AGN within the ALHAMBRA fields, finding a redshift precision of $\sigma \simeq 1\,\%$.

As we stated in the previous section, the properties of the filter system of the survey to be analysed are essential to determine i) the approximate minimum EW of the emission lines that can be detected and ii) the redshift precision that can be achieved. In Fig. 4.3 we show an estimation of the minimum equivalent width (as defined in Eq. 4.3) that can be detected in each ALHAMBRA medium-band, as a function of the source's redshift, the magnitude in the band, and using $\sigma_{\mathrm{line}} = 1$. By definition, the value of $\mathrm{EW}_{\mathrm{min}}$ decreases for bright sources (higher SNR) and at high-z. In addition, the value of $\mathrm{EW}_{\mathrm{min}}$ grows for the reddest bands. This is because the CCD efficiency is lower for the reddest bands (see the overall transmission of the ALHAMBRA filter system in Fig. 4.4).

In Table 4.1 we list the AGN emission lines that are potentiality detectable with at least $1\,\sigma$ precision for i) sources with magnitude $\lesssim 21.5$ in the band where the line falls, and ii) an observed central wavelength, $\lambda_{\mathrm{obs}} = \lambda_c(1+z)$, smaller than 9000 Å. In addition, as the ALHAMBRA bands are contiguous, these lines can be detected

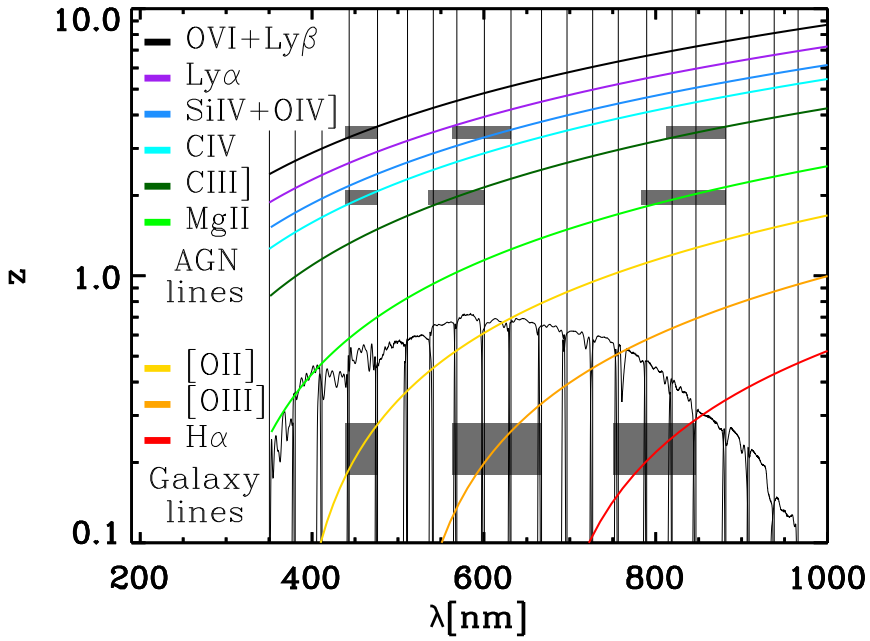


Figure 4.4: Evolution of the central wavelength of AGN and galaxy emission lines as a function of redshift. We also display the transmission curves of the ALHAMBRA medium-band filters, which allow us to see in which band the emission lines are located as a function of z. The grey areas highlight the redshift intervals for which there is a degeneracy among the triplet of galaxy emission lines {[O II], [O III], and H α}, and the triplets of AGN emission lines {C IV, C III], and Mg II} and {O VI+Ly β, Si IV]+O IV], and C III]}.

at any redshift at which they fall within the ALHAMBRA wavelength coverage. We do not look for emission lines with $\lambda_c > 3000$ Å, such as [O II], Hβ, [O III], or H α, because these lines also appear in star-forming galaxies. Whereas it is possible to use them to discriminate between type-I AGN and star-forming galaxies as the lines of type-I AGN are much broader, the low spectral resolution of ALHAMBRA prevent us to securely employ them (we expect this to be possible in surveys with narrower bands). Therefore, given the lines that we can use to detect AGN and their strengths, we will be able to securely identify type-I AGN at $z > 1$ (unobscured broad emission line AGN with no or very little contribution from the host). We thus focus on the detection of type-I AGN in this work, and we tune ELDAR accordingly.

4.3.2 Effects that may reduce the redshift precision and purity

Before optimising ELDAR for detecting type-I AGN in the ALHAMBRA survey, we will explore three effects that may decrease the quality of ELDAR's results: i) confusion between pairs/triplets of AGN and galaxy emission lines, which decreases the purity; ii) confusion between different pairs/triplets of AGN emission lines, which decreases

the redshift precision; and iii) detection of spurious lines, which may decrease both the purity and redshift precision. There are examples of these effects in Appendix A.

The confusion between different pairs/triplets of emission lines arises due to the finite wavelength resolution of medium- and narrow-band surveys. The misidentification appears at redshifts where the relative difference between the central wavelengths of different pairs/triplets of emission lines is the same, and thus they may fall in the same bands. The number and width of these redshift intervals depend on the width of the survey bands, where the narrower the bands the smaller the incidence. In Fig. 4.4 we display the observed central wavelength of several AGN and galaxy emission lines as a function of redshift. Moreover, we plot the transmission curves of the ALHAMBRA medium-band filters. They guide the eye to see the band where emission lines fall as a function of z. There is only one redshift interval for which a triplet of galaxy emission lines can be confused with triplets of AGN emission lines. This is at $z \simeq 0.2$ where the galaxy lines {[O II], [O III], and H α} can be confused with the AGN lines {C IV, C III], and Mg II} at $z \simeq 2$ and {O VI+Ly β, Si IV+O IV], and C III]} at $z \simeq 3.5$. In the figure, these redshift intervals are highlighted in grey. Nevertheless, the incidence of line misidentification between pairs of galaxy and AGN emission lines is much higher. As a consequence, low-z star-forming galaxies may be confused as high-z type-I AGN, which is important because their number density is much higher than the number density of type-I AGN. In addition to galaxies classified as AGN, misidentification of emission lines may lead to catastrophic redshift solutions. We study this in detail in §4.4.2.

We define a spurious line as a line detected by ELDAR in a band where no emission lines should fall according to $z_{\rm spec}$. They appear due to photometric errors and their incidence depends on the criterion chosen to confirm emission lines, $\sigma_{\rm line}$, where the smaller its value the higher the frequency.

To get a rough estimation of the impact of spurious lines in ALHAMBRA, we consider the case of a mock source with a flat SED. Then, we compute the magnitude and uncertainty in each ALHAMBRA medium-band, where the uncertainties are computed using ALHAMBRA empirical errors[2]. After that, we perturb the magnitude in each band 10^5 times using Gaussian distributions with width equal to the $1\,\sigma$ uncertainty in the band, generating 10^5 random realisations of the mock source. In Fig. 4.5 we show two of these realisations. In the first one, ELDAR detects spurious lines in the 2nd and 4th bands for $\sigma_{\rm line} = 1$, the same bands where {C IV and C III]} fall at $z = 1.48$. In the second, ELDAR detects spurious lines in the 4th, 8th, and 12th bands for $\sigma_{\rm line} = 1$, the same bands where {Ly α, C IV, and C III]} at $z = 2.76$. Therefore, these objects could be wrongly classified as type-I AGN by certain configurations of ELDAR.

The number of sources wrongly confirmed as type-I AGN due to spurious lines depends on $\sigma_{\rm line}$ and $\mathcal{N}$, where the higher their values the lower the contamination. Therefore, it is very important to take this into account before choosing the value of $\mathcal{N}$ and $\sigma_{\rm line}$. Moreover, the incidence of spurious confirmations is even higher for objects with real emission lines, as the combination of spurious and real lines may lead to misclassification of sources and/or catastrophic redshift solutions. Finally,

[2]We use all the ALHAMBRA objects with good photometry and F814W > 24.5 to compute empirical error curves for each ALHAMBRA band as a function of the magnitude in the band.

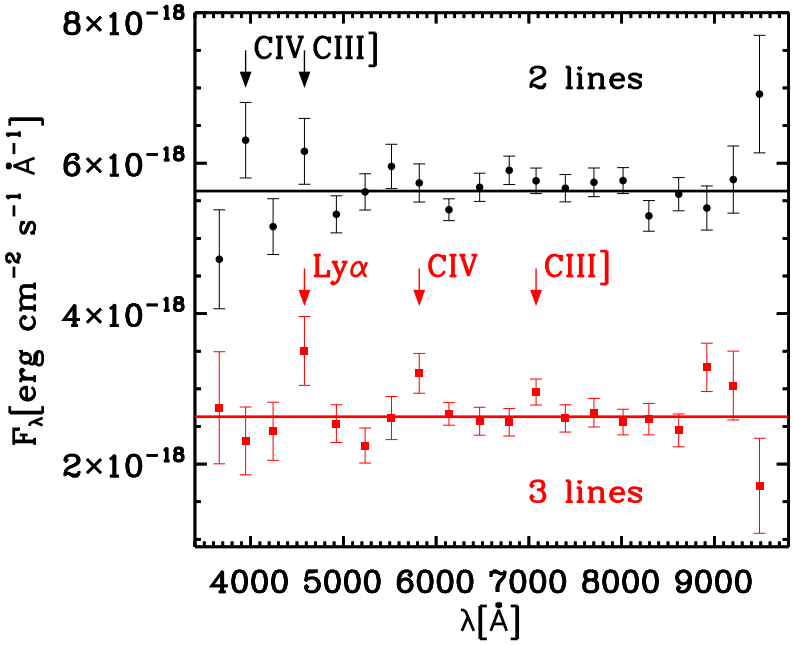


Figure 4.5: Two ALHAMBRA mock realisations of a source with a flat SED and F814W = 21.5. The black points show fluxes measured in each ALHAMBRA medium-band in the first realisation and the red squares in the second. The measured fluxes are not on the top of the solid lines, which indicate the underlying SED of the mock source, due to photometric errors. The fluxes of the first (second) realisation are displaced $+(-)1.5 \times 10^{-18}\,\mathrm{erg\,cm^{-2}s^{-1}\AA^{-1}}$ for visual purposes.

another effect that increases the number of spurious lines is a bad calibration of the zeropoints; however, this is not an issue for as because ALHAMBRA zeropoints are very robust (for a detailed discussion see M14).

4.3.3 Specific configuration of ELDAR for the ALHAMBRA survey

Here we configure ELDAR to identify type-I AGN in the ALHAMBRA survey. In order to do this, we start by optimising LePHARE, and then we tune the spectro-photometric configuration to extract two samples of type-I AGN, where the first prioritises completeness and the second purity.

Given the width of the ALHAMBRA bands, the only AGN that we can securely detect are the ones with strong emission lines, i.e. type-I AGN. Consequently, we will only introduce templates describing the SED of these objects in the extragalactic library of LePHARE. Specifically, we select the empirical templates of quasars and AGN used in Salvato et al. (2009, 2011) and the synthetic templates of quasars included in the LePHARE distribution. The resulting library encompasses 49 templates, where 31 of them are synthetic templates that employ different power laws for the AGN continuum and EWs for the emission lines. From this list, we will select the ones that

produce the best results in terms of completeness and redshift precision for a sample of spectroscopically-known AGN within the ALHAMBRA fields, which we call AGN-S.

The AGN-S sample is obtained by performing a crossmatch between the spectroscopically confirmed point-like type-I AGN (sources with Q or A flags) from the Million Quasar Catalogue (MQC)[3] (Flesch 2015, references within) and the ALHAMBRA sources with F814W < 23. The MQC is a largely complete compendium of AGN from the literature through 21 June 2016. We do the match for objects separated by less than 2 arcsec and, in the two cases where we find two ALHAMBRA sources separated less than 2 arcsec to one object of the MQC, we visually confirm the match by looking at the ALHAMBRA photometry (in both cases we validate the match with blue objects that clearly exhibit broad emission lines). We end up with a total of 220 sources for the AGN-S sample

Table 4.2: Extragalactic templates that we introduce in LePHARE.

Index	Template	Class
1	I22491_70_TQSO1_30	Quasar 30 % + Galaxy 70 %[1]
2	I22491_60_TQSO1_40	Quasar 40 % + Galaxy 60 %[1]
3	I22491_50_TQSO1_50	Quasar 50 % + Galaxy 50 %[1]
4	I22491_40_TQSO1_60	Quasar 60 % + Galaxy 40 %[1]
5	pl_I22491_30_TQSO1_70	Quasar 70 % + Galaxy 30 %[1]
6	pl_I22491_20_TQSO1_80	Quasar 80 % + Galaxy 20 %[1]
7	pl_QSO_DR2_029_t0	Quasar low lum.[1]
8	pl_QSOH	Quasar high lum.[1]
9	pl_TQSO1	Quasar high IR lum.[1]
10	qso-0.2_84	Quasar synthetic[2]
11	QSO_VVDS	Quasar[3]
12	QSO_SDSS	Quasar[4]

References. [1] Salvato et al. (2009), [2] LePHARE distribution, [3] VVDS composite (Gavignaud et al. 2006), and [4] SDSS composite (Vanden Berk et al. 2001). Templates starting with pl_ are extended into the UV using a power law (see Salvato et al. 2009).

Then, to select the final list of templates:

- We start by running LePHARE on the AGN-S sample, and then we reject all the templates that are not assigned to any source at its spectroscopic redshift.
- We compute the redshift precision for the AGN-S sample (using the mode of the PDZ produced by LePHARE) employing all the remaining templates but one at a time. After that, we reject the templates that do not improve the redshift precision.

[3] http://heasarc.gsfc.nasa.gov/w3browse/all/milliquas.html

We end up with the 12 templates listed in Table 4.2 and plotted in Fig. 4.6. The templates 1-6 are from Salvato et al. (2009) and show the SED of a starburst galaxy and a type-I AGN in different proportions; the templates 7-9 are also from Salvato et al. (2009) and present the SEDs of pure type-I AGN; the template 10 is from the LePHARE distribution and describes the SED of a synthetic quasar; finally, the templates 11-12 are quasar composite templates, the first from the VIMOS-VLT Deep Survey (VVDS, Gavignaud et al. 2006) and the second from the SDSS survey (Vanden Berk et al. 2001).

All these templates are at rest-frame. In order to compute precise redshifts for type-I AGN, we have to define the redshift range and redshift step for displacement (see discussion in §4.2.1). We set the maximum redshift to be $z = 6$, as at $z > 6$ most of the AGN emission lines of Table 4.1 are outside the ALHAMBRA medium-band wavelength coverage, making impossible for ELDAR to confirm type-I AGN. As for the redshift step, we set it to be $\Delta z = 0.01$, which is approximately the redshift precision that can be achieved with the ALHAMBRA data. We have checked that a finer redshift step does not produce a higher redshift precision for the AGN-S sample.

We impose a flat prior on the absolute magnitude in the ALHAMBRA band F830W of $-30 < M_{\rm F830W} < -20$, which is a luminosity prior appropriate for our search of type-I AGN. The prior is set in the F830W band because is the medium-band whose central wavelength is the closest to the one of the synthetic detection band of ALHAMBRA, F814W.

After tuning LePHARE, we need to define the configuration of the spectro-photometric confirmation step. We have to select $\mathcal{N}$ and $\sigma_{\rm line}$, whose values depend on the levels of purity and completeness that we want to achieve. In the present analysis we decided to extract two samples of type-I AGN defining two different ELDAR configurations, the first prioritising completeness and the second purity. The specific characteristics of these configurations are the following:

- **2-lines mode**: We require for this mode $\mathcal{N} = 2$, $\sigma_{\rm line} = 1.5$, and F814W = 22.5 as limiting magnitude. The first requirement sets the minimum redshift for confirming sources to $z = 1$, as it is the minimum redshift at which two AGN emission lines of Table 4.1 fall within the ALHAMBRA medium-band wavelength coverage.
- **3-lines mode**: We demand for this configuration $\mathcal{N} = 3$, $\sigma_{\rm line} = 0.75$, and F814W = 23 as limiting magnitude. The demand of detecting at least three AGN emission lines fixes the minimum redshift to $z = 1.5$. It also enables the possibility of detecting fainter sources and lines with lower contrast, as a higher value of $\mathcal{N}$ reduces the galaxy contamination (see Appendix B). Nevertheless, we relax $\mathcal{N} = 3$ to $\mathcal{N} = 2$ for sources at $z_{\rm phot} > 5$ to increase the completeness, as the total number of emission lines within the ALHAMBRA medium-band wavelength coverage at $5 < z < 5.6$ is 3 (O VI+Ly β, Ly α, and Si IV+O IV]) and at $5.6 < z < 6$ is just 2 (O VI+Ly β and Ly α).

These two configurations are selected to minimise the fraction of false detections while pushing the completeness and magnitude limit. In Appendix B we explore the completeness, redshift precision, and galaxy contamination in the case of different values of $\mathcal{N}$, $\sigma_{\rm line}$, and F814W magnitude cuts.

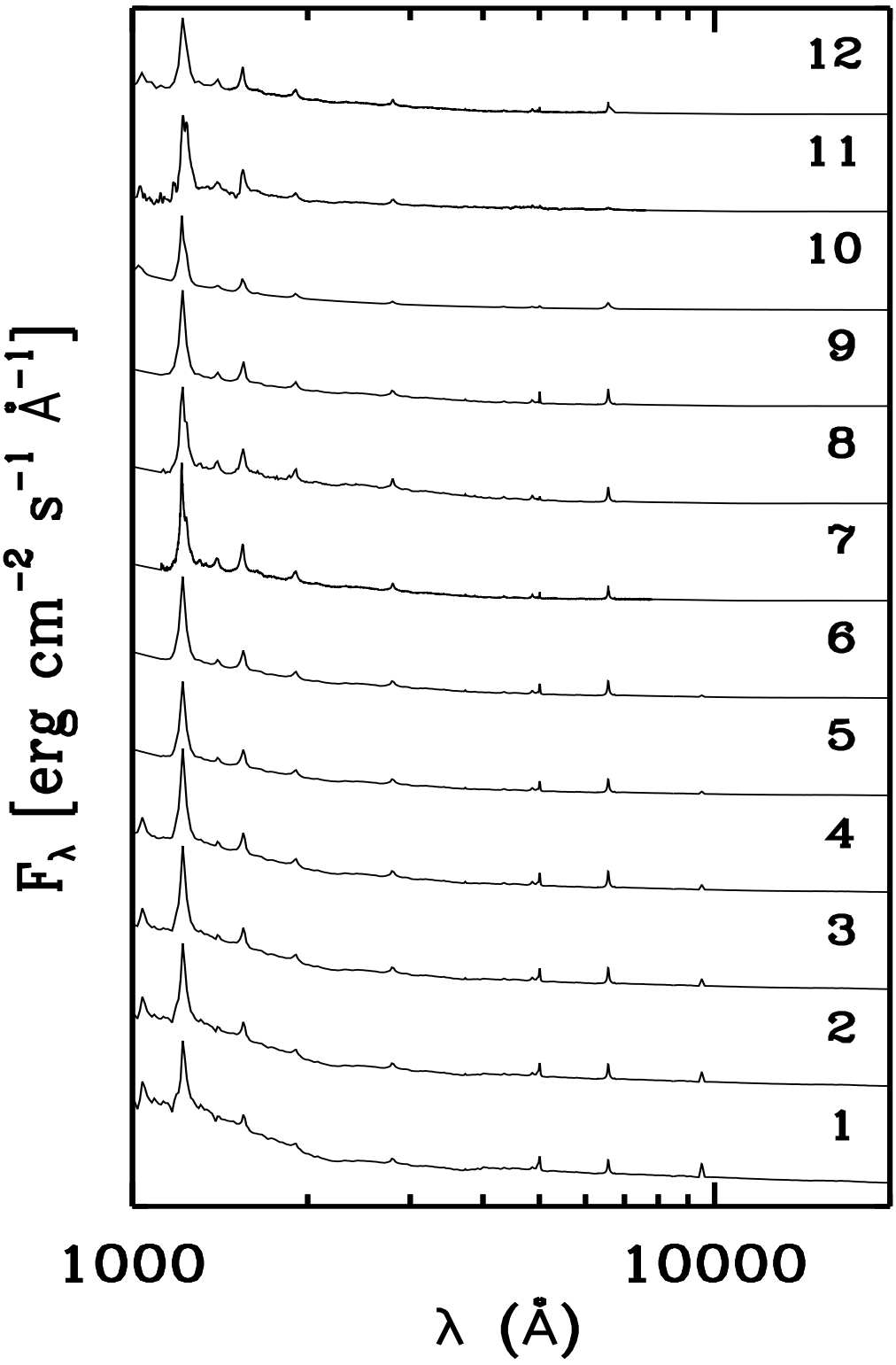


Figure 4.6: Extragalactic templates that we introduce in LePHARE. They are sorted in the same order as in Table 4.2 and the fluxes are expressed per unit wavelength.

We set an additional requirement for objects with the Lyman-break within the ALHAMBRA medium-band wavelength coverage. We demand that these objects cannot have a $3\,\sigma$ flux detection in more than one band with a central wavelength smaller than the Lyman-break (912 Å) at rest-frame. We allow flux detection in one band because of metal lines with $\lambda_c < 912$ Å, such as NeVIII and MgX. This criterion aims at rejecting low-z galaxies for which the 4000 Å break is confused with the Lyman-break.

Finally, as low-z galaxies have extended Point Spread Function (PSF) whereas type-I AGN at $z > 1$ are point-like, we will not apply ELDAR to sources with extended morphology. To characterise the morphology, we will employ the Stellarity parameter of SExtractor (Bertin & Arnouts 1996), which is one for point-like sources and zero for extended ones, and we will not run ELDAR on sources with Stellarity < 0.2. We do not select a higher cut-off because in ground base surveys, if data obtained with bad seeing are stacked together, the PSF gets smeared (see Hsu et al. 2014, for a demonstration with AGN). However, if the value of Stellarity is smaller than 0.2, the probability of the source to be point-like is very low for ALHAMBRA sources with F814W < 23 (see M14). We explore further contamination from low-z galaxies in §4.4.2.

The same steps followed here to tune ELDAR for the ALHAMBRA survey can be used to adjust the ELDAR configuration for surveys with different filter systems and depths.

4.3.4 Summary of the ELDAR configuration for the ALHAMBRA survey

In §4.2 we described the main characteristics of ELDAR and in §4.3.3 we tuned our methodology to identify type-I AGN with the ALHAMBRA data. In what follows, we summarise how ELDAR works and its main properties for this specific case:

- We start by running LePHARE over all non-extended sources of the ALHAMBRA survey (Stellarity > 0.2) using templates describing the SEDs of stars and type-I AGN. We then reject the objects best-fitted by stellar templates.
- We look for the AGN emission lines gathered in Table 4.1 at the redshifts in which the value of the PDZ is greater than 0.5. We then confirm AGN emission lines detected with $\sigma_{\rm line} = 1.5$ for the 2-lines mode and with $\sigma_{\rm line} = 0.75$ for the 3-lines mode. This requirement sets a minimum redshift for confirming sources of $z_{\rm min} = 1$ for the former and $z_{\rm min} = 1.5$ for the latter.
- For the 2-lines mode, we confirm the objects with F814W < 22.5 and at least 2 detected AGN emission lines. For the 3-lines mode, we validate the sources with F814W < 23 and at least 3 detected AGN emission lines. Additionaly, we require the detection of Ly α for objects at $z_{\rm phot} > 2$ and that the flux in the band where Ly α falls has to be greater than the 75 % of the maximum flux in any of the other bands. We also demand no flux detection in more than one band whose central wavelength is smaller than the Lyman-break at rest-frame.
- Finally, we compute the redshift of the confirmed sources using Eq. 4.5 (see the mock example in Fig. 4.2).

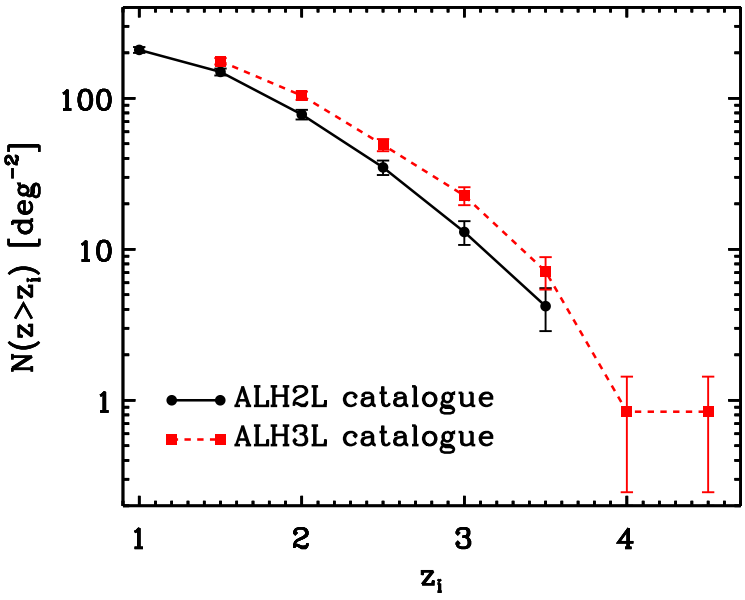


Figure 4.7: Number density of type-I AGN at $z > z_i$. The black solid line indicates the results for the ALH2L catalogue and the red dashed line for the ALH3L catalogue, which have a limiting magnitude of F814W = 23 and F814W = 22.5, respectively. Note that we only include the objects within the ALHAMBRA mask and that the error bars denote Poisson errors.

4.4 The ALHAMBRA type-I AGN catalogues

To prove the effectiveness of ELDAR and to characterise its properties, in this section we apply the 2- and 3-lines modes of ELDAR to the ALHAMBRA data. We will end up with two type-I AGN samples, the ALH2L an ALH3L catalogues, respectively. Then, we will present their properties and we will discuss their quality in terms of redshift precision, completeness, and galaxy contamination.

We start by selecting the ALHAMBRA sources to be analysed. From the 446 361 sources of the M14 catalogue with good photometry (Satur_Flag and DupliDet_Flag equal to zero), we pick 41 367 no extended objects (Stellarity > 0.2) with F814W < 23. We then run LePHARE on these sources, rejecting the 20 580 objects best-fitted by stellar templates. After that, we apply the 2- and 3-lines modes of ELDAR to the remaining sample, ending up with 585 type-I AGN with $z > 1$ and F814W < 22.5 (ALH2L catalogue), and 494 type-I AGN with $z > 1.5$ and F814W < 22.5 (ALH3L catalogue), respectively. We note that 461 sources of the first and 408 of the second are not spectroscopically-known. Both catalogues are publicly available and they are detailed in Appendix C.

To compute the number density of type-I AGN for the ALH2L and ALH3L catalogues, we need the effective area of the ALHAMBRA survey. To obtain it, we employ a mask generated by Arnalte-Mur et al. (2014), which excludes low exposure time areas, obvious defects in the images, and circular regions around saturated stars. After

applying this mask, the effective area of the ALHAMBRA survey is 2.381 deg^2. We apply the same mask to the ALH2L and ALH3L catalogues, finding 498 and 419 objects within the mask, respectively. Using these numbers, we obtain a surface number density of $\simeq 209\,\mathrm{deg}^{-2}$ for the ALH2L catalogue and $\simeq 176\,\mathrm{deg}^{-2}$ for the ALH3L catalogue. In Fig. 4.7 we show the number density for the two catalogues as a function of redshift.

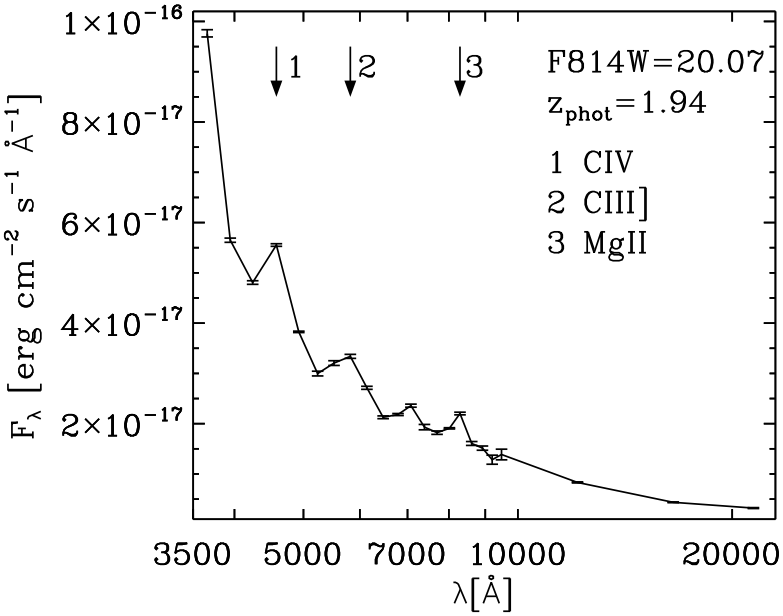


Figure 4.8: Object of both the ALH2L and ALH3L catalogues at $z_{\mathrm{phot}} = 1.935$ with identification number (ID) ALH2L346 and ALH3L186, respectively. Arrows point to the bands where AGN emission lines are confirmed.

In Figs. 4.8-4.10 we show examples of objects from the ALH2L and ALH3L catalogues. In the figures we employ arrows to point to the bands where ELDAR detects AGN emission lines. In Fig. 4.8 we display an object at $z_{\mathrm{phot}} = 1.935$ belonging to both the ALH2L and ALH3L catalogues, for which ELDAR detects the lines C IV, C III], and Mg II. Despite the very blue and steep continuum, our methodology, which assumes a flat continuum, is able to clearly detect all the lines that it is looking for. This source is best-fitted by the qso-0.2_84 template with a very low colour excess ($E(B-V) = 0.02$).

In Fig. 4.9 we present an object of both the ALH2L and ALH3L catalogues at $z_{\mathrm{phot}} = 3.258$ for which ELDAR detects the complex O VI + Ly β and the lines Ly α, C IV, and C III]. The AGN continuum is approximately flat at this redshift, and thus the detection of AGN emission lines is straightforward. It best-fitted by the qso-0.2_84 template without any extinction.

In Fig. 4.10 we display the SED of an object of the ALH3L catalogue at $z_{\mathrm{phot}} = 4.549$ for which ELDAR detects the complexes O VI + Ly β and Si IV + O IV], and the lines Ly α and C IV. It is not included in the ALH2L catalogue because its magnitude, F814W = 22.65, is dimmer than the magnitude limit for this catalogue,

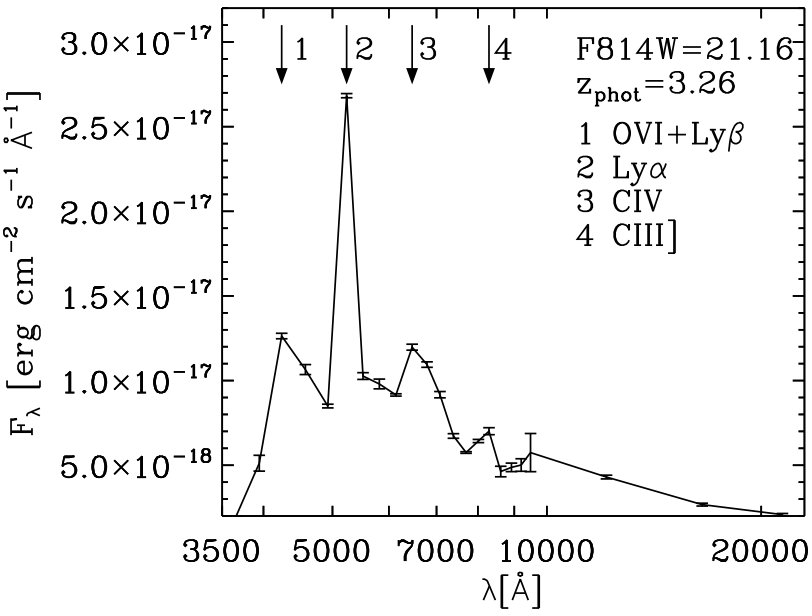


Figure 4.9: Object of the ALH2L and ALH3L catalogues at $z_{\rm phot} = 3.258$ with ID ALH2L560 and ALH3L450, respectively.

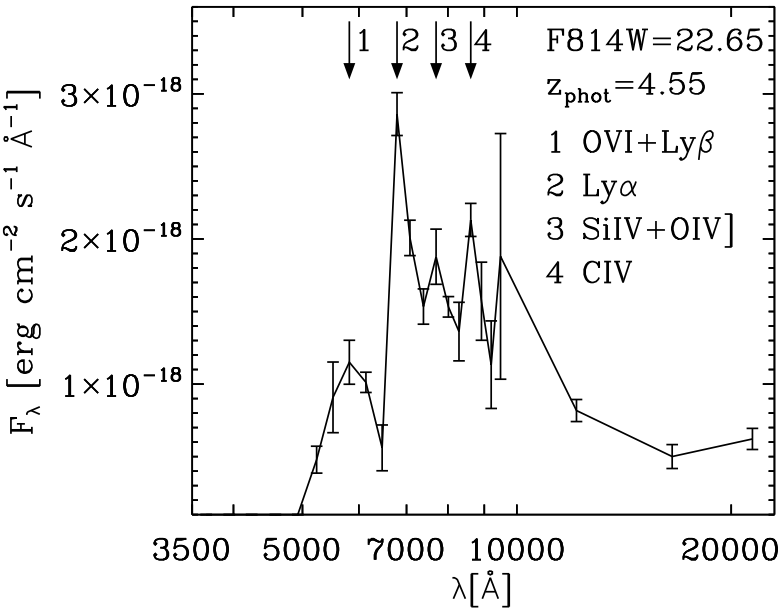


Figure 4.10: Object of the ALH3L catalogue at $z_{\rm phot} = 4.549$ with ID ALH3L490.

F814W = 22.5. This object is best-fitted by the qso-0.2_84 template with a very low colour excess ($E(B-V) = 0.04$). Moreover, it is one of the eight objects of the ALH3L catalogue at $z_{\rm phot} > 4$, where the one at the highest redshift, $z_{\rm phot} = 5.413$, was spectroscopically confirmed by Matute et al. (2013) at $z_{\rm spec} = 5.410$. The rest of them will be spectroscopically analysed in Chaves-Montero et al. (in prep.).

4.4.1 Properties of the ALH2L and ALH3L catalogues

In this section we show the magnitudes, redshifts, best-fitting templates, and colours of the objects of the ALH2L and ALH3L catalogues.

Table 4.3: ALHAMBRA medium-bands whose central wavelength, λ_c, is the closest to the central wavelength of SDSS broad-bands.

SDSS	$\lambda_{\rm c}$(nm)	ALHAMBRA	$\lambda_{\rm c}$(nm)
u	354	F365W	365
g	477	F489W	489
r	623	F613W	613
i	762	F768W	768
z	913	F923W	923

In the top and bottom panel of Fig. 4.11, we present the magnitude and redshift distribution for the ALH2L and ALH3L catalogues, respectively. As there are not obvious gaps in the redshift distribution of these catalogues, we conclude that ELDAR uniformly identifies type-I AGN as a function of redshift. This is thanks to the continuity of the ALHAMBRA medium-bands. Non-contiguous bands would introduce gaps in the redshift distribution due to emission lines falling in between them.

In Fig. 4.12 we display the magnitude and best-fitting template distribution for the ALH2L and ALH3L catalogues in the top and bottom panels, respectively. The most selected best-fitting template for both catalogues describes the SED of a high luminosity quasar (template number 8, pl_QSOH) and the second most selected template depicts the SED of a synthetic quasar whose continuum emission follows a power law (template number 10, qso-0.2_84). In addition, the most selected template depends on the redshift of the source, it is pl_QSOH for sources at $z < 2$ and qso-0.2_84 for objects at $z > 2$. We find that the 85 % of the sources of the ALH2L and ALH3L catalogues have a low extinction ($E(B-V) < 0.2$), in agreement with the fact that these objects are unobscured type-I AGN.

To investigate whether the colour locus of the objects identified by ELDAR is the same as the colour locus of quasars targeted by traditional colour selections, we compare the colour locus of the ALH2L and ALH3L catalogues with the colour locus of SDSS quasars.

The SDSS photometric filter system includes five broad-bands (u, g, r, i, z) while the ALHAMBRA survey comprises twenty medium-bands. To perform a comparison between SDSS colours and ALHAMBRA colours, we match each SDSS broad-band to the ALHAMBRA medium-band with the closest central wavelength (see Table 4.3).

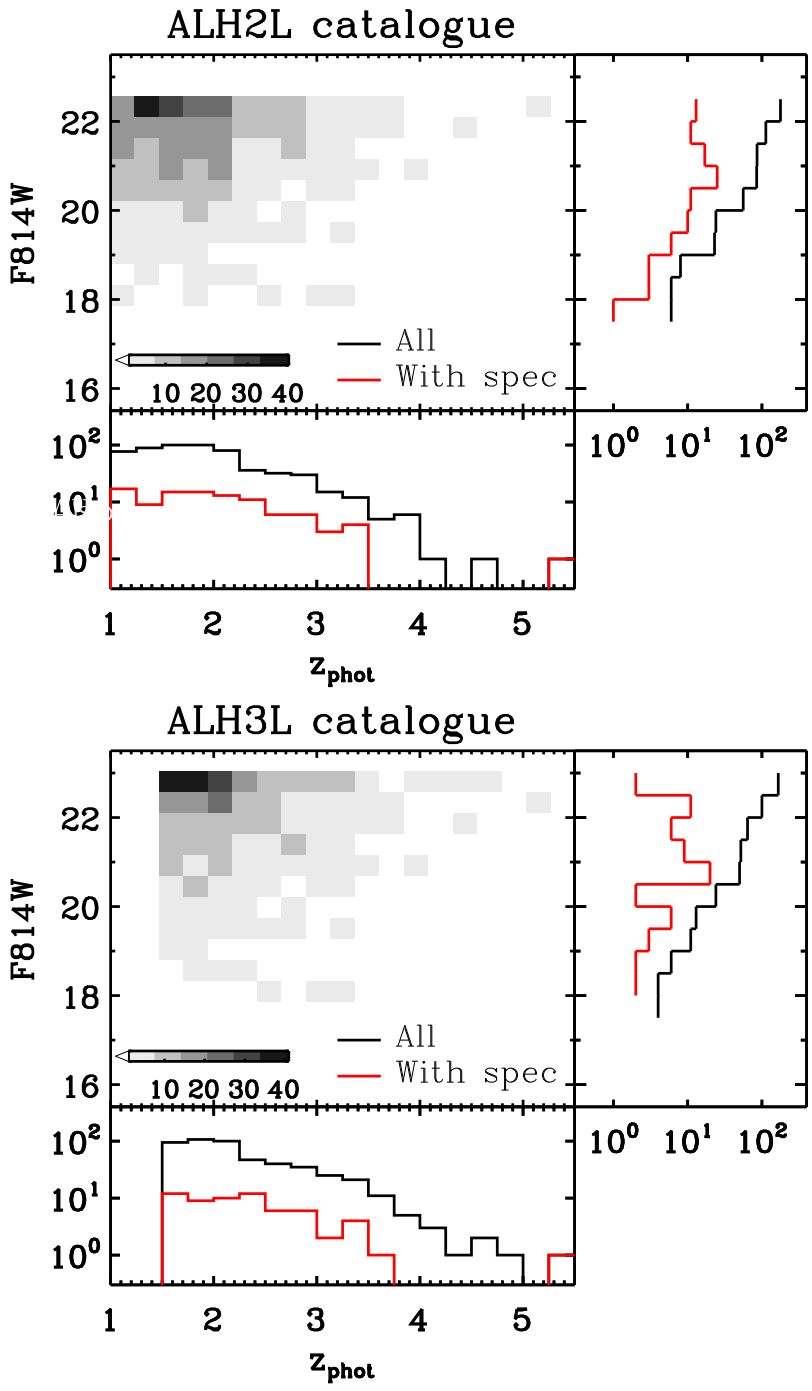


Figure 4.11: photo-z and F814W magnitude distribution for the ALH2L (top panel) and ALH3L (bottom panel) catalogues. The black histograms denote all the objects of the catalogue and the red ones sources spectroscopically-known.

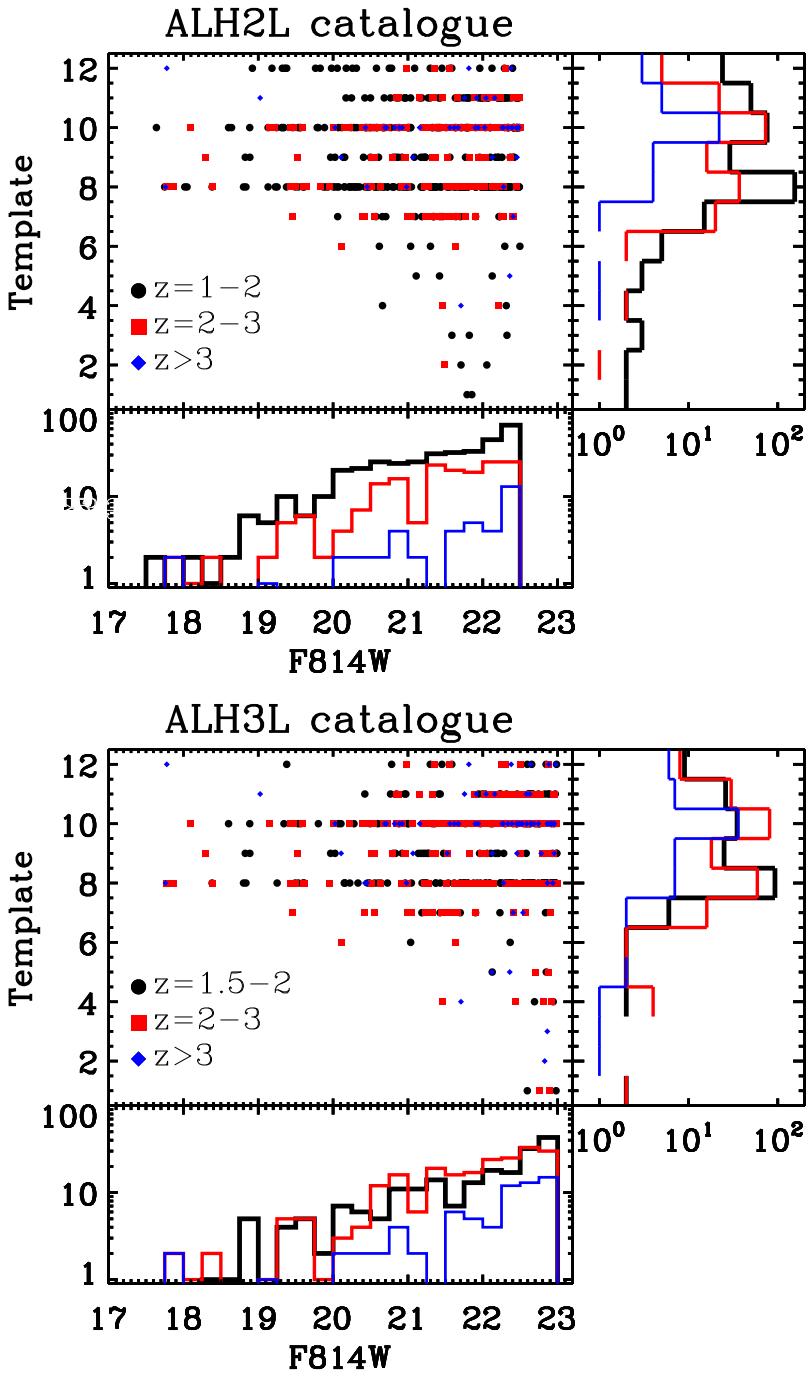


Figure 4.12: Frequency of best-fit template and F814W magnitude distribution for the ALH2L (top panel) and ALH3L (bottom panel) catalogues. The colour and shape of the points indicate the redshift, as shown by the legend. (The SEDs of all of the templates is shown in Fig. 4.6.

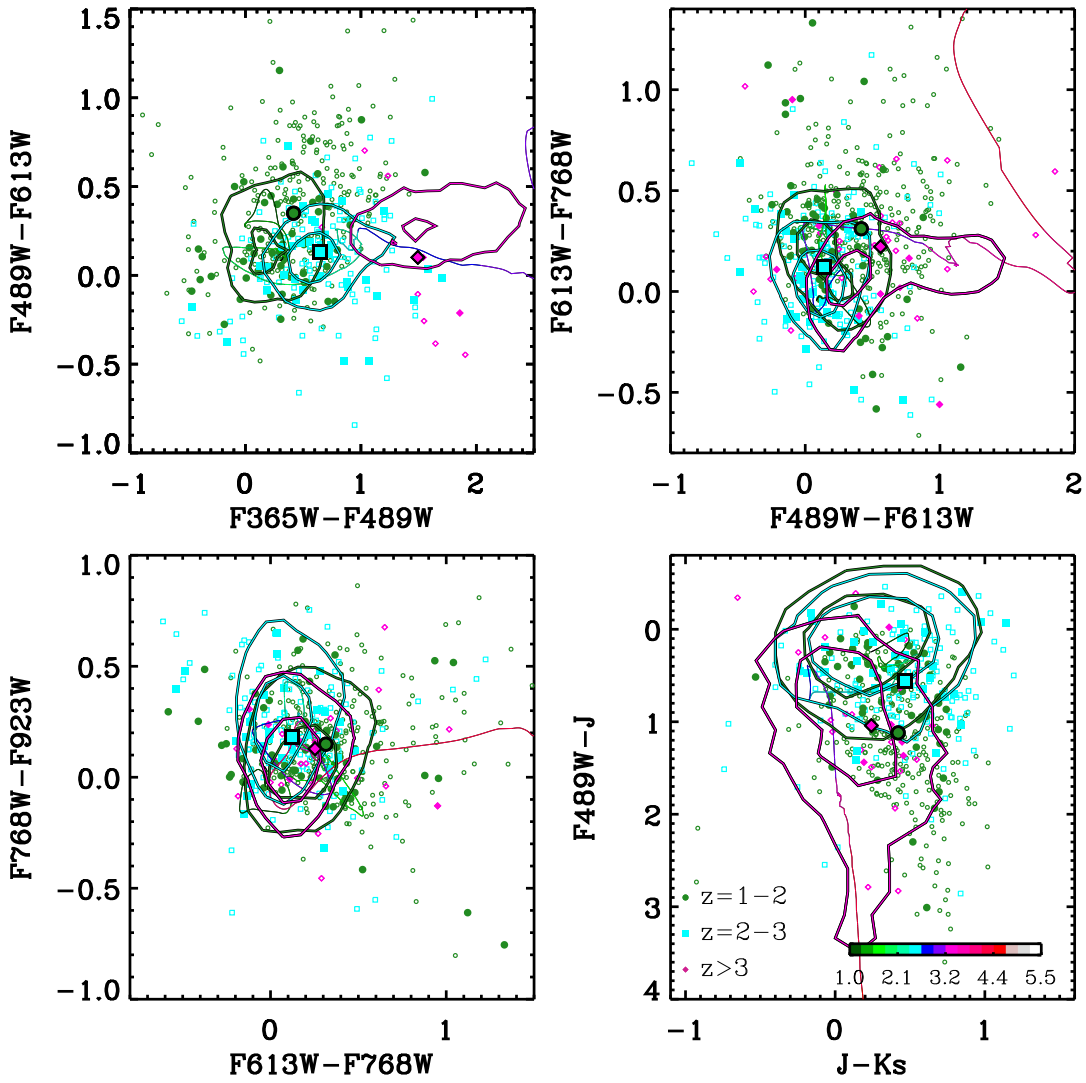


Figure 4.13: Colour-colour diagrams for the ALH2L catalogue. Only objects with photometric errors smaller than 0.2 mag in the bands shown are used to generate each panel. The colour of the symbols and lines indicates the redshift, as stated in the legend. Filled (open) symbols denote ALH2L objects that are (not) in common with the AGN-S sample and big symbols indicate the median colours for all the ALH2L sources at a certain redshift. Contours outline the colour locus of quasars from the SDSS-Data Release (DR)12 Quasar catalogue (top-left, top-right, and bottom-left panels) and the SDSS-DR6 Quasar catalogue with a counterpart in the United Kingdom Infrared Telescope Infrared Deep Sky Survey (UKIDSS)-Large Area Survey (LAS) (bottom right panel), where the inner contour encloses the 0.5 % of the sample and the outer contour the 3 %. The equivalence between ALHAMBRA and SDSS bands is given in Table 4.3. Narrow lines display the evolution of the colours of the template pl_QSOH as a function of z.

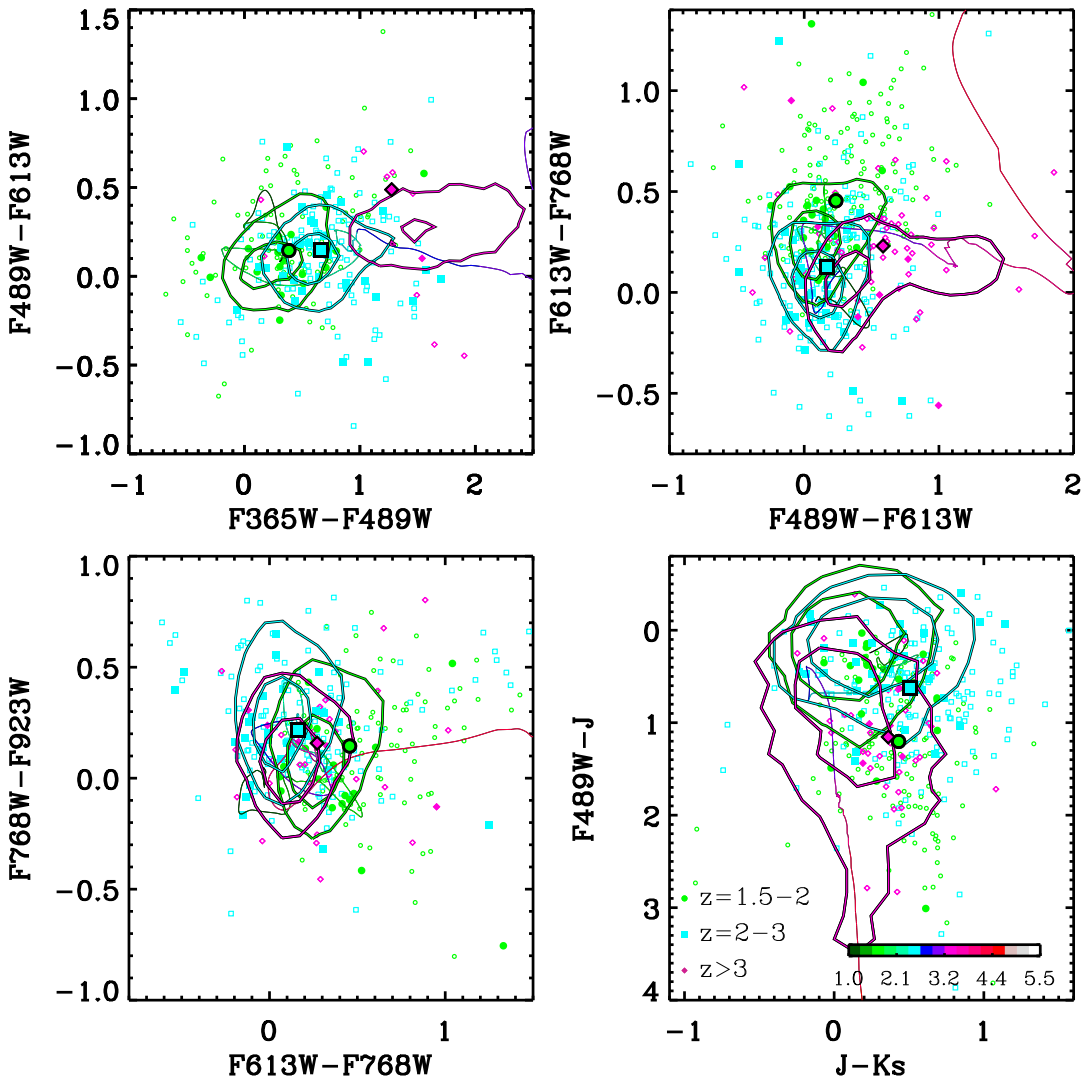


Figure 4.14: Colour-colour diagrams for the ALH3L catalogue. We employ the same colour coding as in Fig. 4.13.

In Fig. 4.13 and 4.14 we display four colour-colour diagrams for the sources of the ALH2L and ALH3L catalogues, respectively. Symbols indicate the colours of individual ALHAMBRA sources and the contours denote the colour loci of spectroscopically confirmed quasars from the SDSS-DR12 Quasar catalogue (Pâris et al. 2017) (top-left, top-right, and bottom-left panels) and the SDSS-DR6 Quasar catalogue with a counterpart in the UKIDSS-LAS (Peth et al. 2011) (bottom right panel). Narrow-lines show the colours of the pl_QSOH template as a function of z. Since ALHAMBRA bands are narrower than SDSS bands, ALHAMBRA colours are more sensitive to emission lines and other features than SDSS colours. The agreement between the colour loci of ALH2L, ALH3L, and SDSS objects is good at high-z; however, at low-z it is smaller because SDSS does not systematically target sources with $z < 2.15$.

4.4.2 Quality of the ALH2L and ALH3L catalogues

In order to asses the quality of the ALH2L and ALH3L catalogues, we need samples of spectroscopically-known type-I AGN and galaxies within the ALHAMBRA fields. We will employ the AGN-S sample (see §4.3.3) and two new samples, the first consists of X-ray selected type-I AGN in the ALHAMBRA COSMOS field and the second includes randomly selected galaxies within same ALHAMBRA field. We name them AGN-X and GAL-S, respectively.

To obtain the AGN-X sample, we perform a crossmatch between the ALHAMBRA sources with F814W $<$ 23 and the 637 type-I AGN from the *Chandra* COSMOS-Legacy X-ray catalog (C-COSMOS) (Civano et al. 2016; Marchesi et al. 2016) with an optical counterpart and spectroscopic redshift. We employ the C-COSMOS catalogue because X-ray selection produces complete samples of type-I AGN (Brandt & Hasinger 2005) and we will use the AGN-X sample to estimate the completeness for the ALH2L and ALH3L catalogues. In addition, X-ray AGN catalogues have a low contamination from galaxies and stars (Lehmer et al. 2012; Stern et al. 2012). Following the same matching procedure as for the AGN-S sample, we end up with a total of 105 sources, where 30 of them are in common with the AGN-S sample.

To get the GAL-S sample we match the objects from the DR2 of the zCOSMOS 10k-bright spectroscopic sample (zCOSMOS) (Lilly et al. 2009) with secure redshift (flags 3.x and 4.x) and the ALHAMBRA sources with F814W $<$ 23. The zCOSMOS includes randomly selected galaxies with F814W $<$ 22.5 at $z_{\rm spec} < 1.5$ in the COSMOS field, where the sampling rate is $\simeq 0.35$ in the ALHAMBRA COSMOS field. Following the same procedure as for the AGN-S sample to do the match, we find a total of 1051 sources.

In Fig. 4.15 we display the magnitude and redshift distribution for the objects of the GAL-S, AGN-S, and AGN-X samples. In the following sections we will employ them to explore, respectively, the galaxy contamination, redshift precision, and completeness produced by the 2- and 3-lines modes of ELDAR. We gather the results for these three samples in Table 4.4.

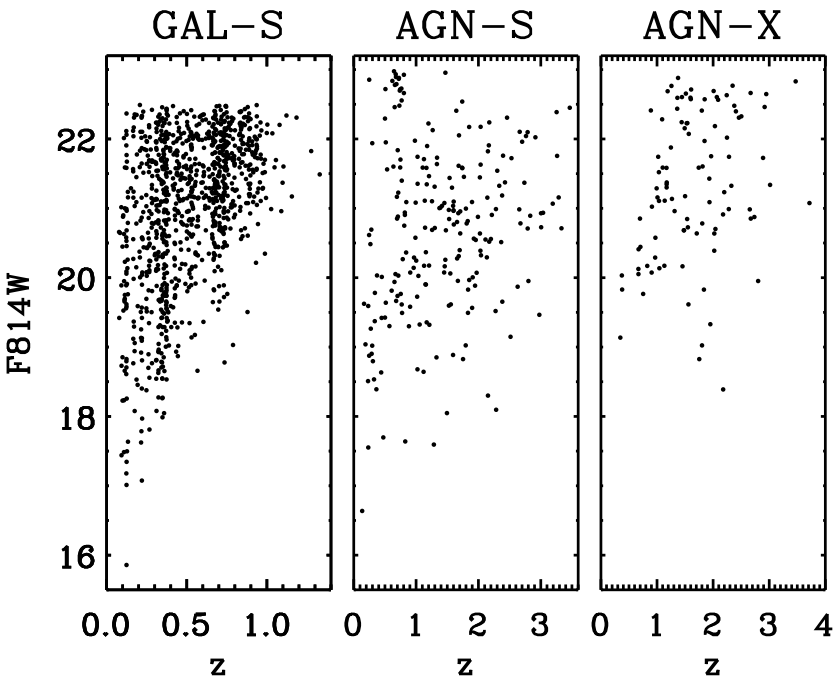


Figure 4.15: Redshift and F814W magnitude distribution for the GAL-S, AGN-S and AGN-X samples. We employ these samples to asses, respectively, the galaxy contamination, redshift precision, and completeness of the ALH2L and ALH3L catalogues.

Redshift precision

We define fraction of redshift outliers in a sample, η, as the percentage of objects with catastrophic redshift solutions for which $|z_{\rm spec} - z_{\rm phot}| > 0.15\,(1 + z_{\rm spec})$. We estimate the fraction of outliers for the ALH2L and ALH3L catalogues by applying the 2- and 3-lines modes of ELDAR to the AGN-S sample, respectively. We find the same fraction of outliers for both catalogues, $\eta = 8.1\,\%$. They are produced due to a degeneracy between pairs of AGN emission lines, such as {C III] and Mg II]} at $z = 1.2$ and {Ly α and C III]} at $z = 2.3$, and {C IV and C III]} at $z = 1.7$ and {Ly α and C IV} at $z = 2.4$. We show the ALHAMBRA photometric data of some of these outliers in Appendix A.

To compute the redshift precision, we employ the normalised median absolute deviation, $\sigma_{\rm NMAD}$, defined by Hoaglin et al. (1983) as

$$\sigma_{\rm NMAD} = 1.48\ \text{median}\left(\frac{|z_{\rm phot} - z_{\rm spec}|}{1 + z_{\rm spec}}\right). \tag{4.6}$$

We employ $\sigma_{\rm NMAD}$ because it is designed to be less sensitive to redshift outliers than the standard deviation of photometric and spectroscopic redshifts. In a distribution without redshift outliers they would have the same value. Applying the 2-lines mode to the AGN-S sample, we obtain a redshift precision of $\sigma_{\rm NMAD} = 0.97$ and using the 3-lines mode, we get $\sigma_{\rm NMAD} = 0.84$. Therefore, the precision reached for type-I AGN using the 3-lines mode of ELDAR is even greater that the one achieved for galaxies

Table 4.4: Results for the AGN-S, AGN-X, and GAL-S samples using the ELDAR's 2- and 3-lines modes.

Sample	Mode	Compl. (%)	$\sigma_{\rm NMAD}$(%)	η(%)
AGN-S	2-lines	72.8	0.97	8.1
	3-lines	64.6	0.84	8.1
AGN-X	2-lines	73.3	1.15	6.8
	3-lines	66.7	0.91	0.0
Sample	Mode	Galaxies confirmed as AGN		
GAL-S	2-lines	4 (31 %)		
	3-lines	1 (9 %)		

Notes. Underlined numbers indicate the estimated redshift precision, completeness, and galaxy contamination for the ALH2L and ALH3L catalogues. The galaxy contamination is extrapolated from the results for the GAL-S sample assuming that the ALHAMBRA COSMOS field is representative for all the ALHAMBRA fields.

and type-I AGN in other ALHAMBRA studies (see M14 and Matute et al. (2012), respectively).

In Fig. 4.16 we show the comparison between the spectroscopic and photo-zs of the sources of the AGN-S sample, where the photo-zs are computed using the 2- and 3-lines modes of ELDAR. The results are similar for the two modes and the main difference between them is that the 3-lines mode only produces outliers that overestimate the redshift of the source whereas the 2-lines mode also generate outliers that underestimate the redshift of the object. In the bottom panel of Fig. 4.16 we display $(z_{\rm spec} - z_{\rm spec})/(1 + z_{\rm spec})$ for the objects of the AGN-S sample, which is a measurement of the photometric accuracy for each source, where we find that it is higher than 3 % for 88 % and 92 % of the sources using the 2- and 3-lines mode, respectively. Therefore, ELDAR produces accurate photo-zs for most of the sources.

In Fig. 4.17 we show the same comparison as in Fig. 4.16 for the sources of the AGN-X sample. The redshift precision that we reach for this sample is compatible with the one reached for the AGN-S sample, whereas the fraction of outliers is smaller. As a consequence, the redshift precision is similar for X-ray and colour selected type-I AGN (a large fraction of the objects from the MQC are selected using colour selections).

In Table 4.5 we gather the redshift precision and outlier fraction for several X-rays selected samples (references in the caption). Most of the sources of the Cardamone et al. (2010); Luo et al. (2010); Hsu et al. (2014) samples are AGN whose SED is dominated by the host galaxy, and thus their photo-z is straightforward to compute because the 4000 Å break is visible. On the other hand, the Salvato et al. (2009, 2011); Fotopoulou et al. (2012); Matute et al. (2012) samples only contains type-I AGN. All these surveys but the Lockman Hole area, which only has broad-band filters, have broad-, medium-, and narrow-band filters. As a consequence, the Lockman Hole sample is the one with the lowest redshift precision and the highest fraction of outliers. In this work, using the AGN-S sample, we obtain the best results

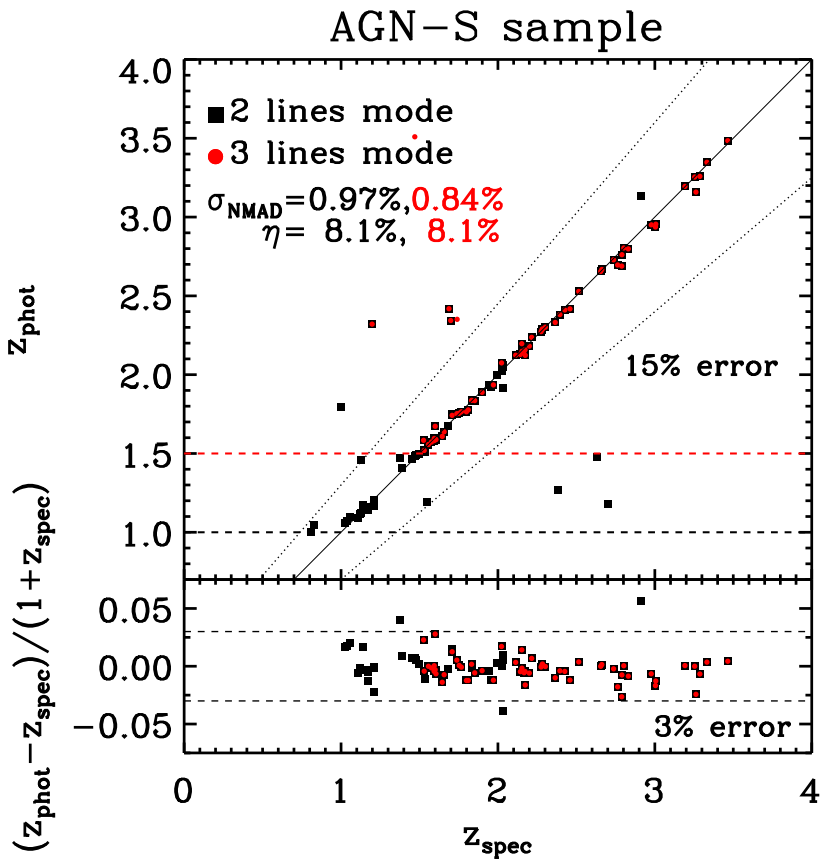


Figure 4.16: Comparison between photometric and spectroscopic redshifts for the AGN-S sample using the 2- and 3-lines modes of ELDAR. Black squares denote objects identified by the 2-lines mode and red circles by the 3-lines mode. The solid line indicates the $1:1$ relation, dotted lines the threshold between good redshift solutions and outliers, and the black (red) dashed line the redshift cut-off for the 2-lines (3-lines) mode. The normalised median absolute deviation, $\sigma_{\rm NMAD}$, and the fraction of outliers, η, are in black for the 2-lines mode and in red for the 3-lines mode. The bottom panel shows a measurement of the photo-z accuracy for each source. Dashed lines indicate 3 % errors.

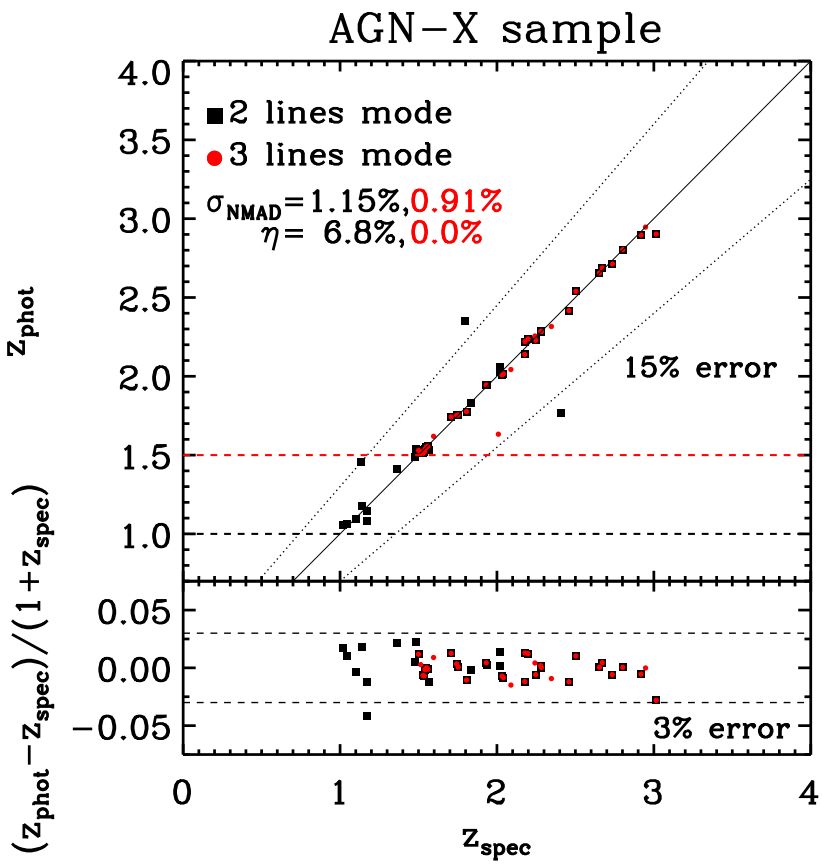


Figure 4.17: Comparison between photometric and spectroscopic redshifts for the AGN-X sample using the 2- and 3-lines modes. We employ the same colour coding as in Fig. 4.16.

in terms of redshift precision, which is because of the contiguous coverage of the optical range by the 20 medium-band filters of ALHAMBRA. Although the fraction of outliers that we obtain is not the lowest one, we want to highlight that the AGN-S sample is not X-ray selected. If we apply our methodology to the AGN-X sample, we find no outliers using the 3-lines mode.

Table 4.5: Redshift precision and fraction of outliers for different AGN/quasar catalogues.

Ref.	Bands	Depth	$\sigma_{\rm NMAD}(\%)$	$\eta(\%)$
(a)	30	$i^*_{AB} < 22.5$	1.2	6.3
(b)	32	$R < 26$	1.2	12.0
(c)	42	$R < 26$	5.9	8.6
(d)	31	$i^*_{AB} < 22.5$	1.1	5.1
(e)	21	$R_c < 22.5$	8.4	21.4
(f)	23	$m_{678} < 23.5$	0.9	12.3
(g)	50	$R < 23$	1.1	4.2
(h)	23	F814W < 22.5	0.97	8.1
(i)	23	F814W < 23	0.84	8.1

Notes. (a) X-ray Multi-Mirror Mission (XMM-Newton)-COSMOS (QSOV sample, Salvato et al. 2009). (b) The Multiwavelength Survey by Yale-Chile (X-ray sources, Cardamone et al. 2010). (c) *Chandra* Deep Field-South (X-ray sources, Luo et al. 2010). (d) XMM-Newton- and *Chandra*-COSMOS (QSOV sample, Salvato et al. 2011). (e) Lockman Hole area (QSOV sample, Fotopoulou et al. 2012) (f) ALHAMBRA (Matute et al. 2012). (g) Extended *Chandra* Deep Field South (X-ray sources, Hsu et al. 2014). (h) ALH2L catalogue (this work). (i) ALH3L catalogue (this work).

Purity

Because of their large number density and emission lines, star-forming galaxies are potentially the largest sample of objects that may be incorrectly classified as type-I AGN by ELDAR. We estimate the galaxy contamination in the ALH2L and ALH3L catalogues by applying the 2- and 3-lines modes to the GAL-S sample. This sample allows us to estimate the galaxy contamination up to F814W = 22.5.

After applying the 2- and 3-lines modes to the 1051 galaxies of the GAL-S sample, we end up with a total of 4 and 1 objects wrongly classified as type-I AGN, respectively. All of them show clear emission lines, have values of Stellarity > 0.6, and are at $z_{\rm spec} < 0.35$. Therefore, they are low-z star-forming galaxies.

In Fig. 4.18 we show the only galaxy of the GAL-S sample that it is wrongly classified as type-I AGN by both the 2- and 3-lines modes. It is a point-like object (Stellarity = 0.95) at $z_{\rm spec} = 0.17$. This source is confirmed by our methodology because there is a degeneracy between the triplet {[O II], [O III], and H α} at $z = 0.17$ and the triplet {C IV, C III], and Mg II} at $z = 1.76$.

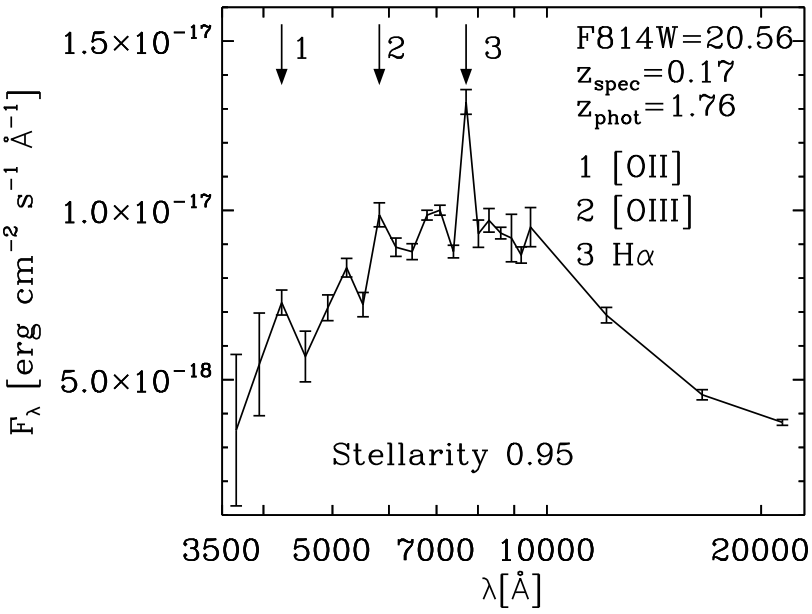


Figure 4.18: Object of the GAL-S sample at $z_{\rm spec} = 0.17$ that it is classified as type-I AGN by the 2- and 3-lines modes at $z_{\rm phot} = 1.76$. This source is confirmed by our methodology because there is a degeneracy between the triplet {[O II], [O III], and H α} at $z = 0.17$ and the triplet {C IV, C III], and Mg II} at $z = 1.76$.

The other galaxies that are wrongly classified at type-I AGN by the 2-lines mode are objects for which there is a degeneracy between pairs of galaxy emission lines and pairs of AGN emission lines. None of them is confirmed due to spurious lines.

The effective area of the ALHAMBRA COSMOS field is $0.203\,{\rm deg}^2$, which is 8.5 % of the total effective area of the ALHAMBRA survey, $2.381\,{\rm deg}^2$. To compute the galaxy contamination for the ALH2L and ALH3L catalogues, we will assume that the ALHAMBRA COSMOS field is representative for the rest of the ALHAMBRA fields. As the sampling rate for zCOSMOS is $\simeq 0.35$ within the ALHAMBRA COSMOS field and 87 % of the galaxies at $z_{\rm spec} < 0.35$ has secure redshifts, we estimate a galaxy contamination of 154 objects for the ALH2L catalogue and 38 for the ALH3L catalogue. This corresponds to a galaxy contamination of 31 % for the first and 9 % for the second. On the other hand, ELDAR assigns photo-z smaller than $z_{\rm phot} = 2.1$ to the 4 galaxies wrongly classified as type-I AGN, and thus we do not expect any galaxy contamination at $z > 2.1$.

We do not explore the contamination from stars because we reject all the sources best-fitted by stellar templates. It is possible that stellar types with a very blue SED, e.g. O, A, and B could be best-fitted by AGN templates; however, they would be rejected during the spectro-photometric step because they do not present emission lines with EWs large enough to be detected in ALHAMBRA. Another source of contamination could be Wolf-Rayet stars since they present broad emission lines of ionized helium, carbon, and nitrogen. Nevertheless, the predicted number density of Wolf-Rayet stars in our quadrant of the galaxies is smaller than 1600 (van der Hucht

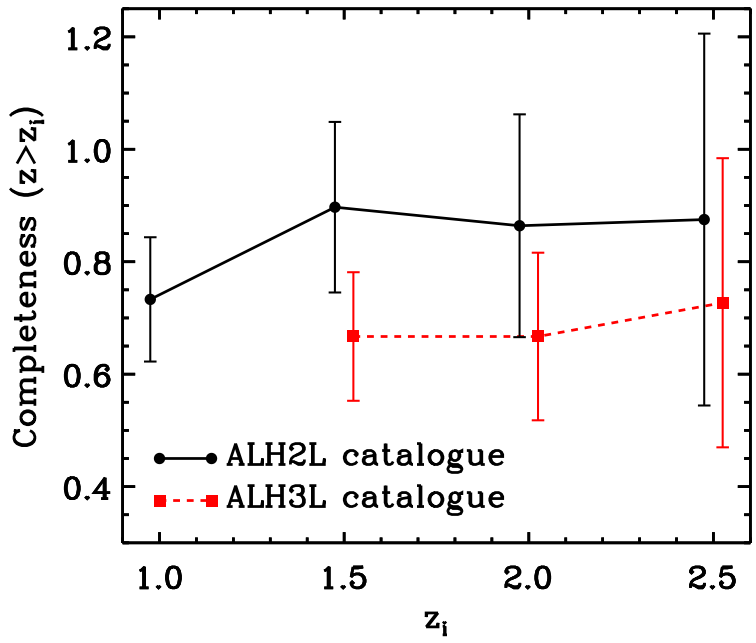


Figure 4.19: Completeness at $z > z_i$ for the ALH2L and ALH3L catalogues. The completeness is estimated using the AGN-X sample.

2001), and thus they cannot be an important source of contamination.

Completeness

To estimate the completeness of the ALH2L and ALH3L catalogues, we apply the 2- and 3-line modes to the AGN-X sample. We employ this sample because, as we noticed before, X-ray selection produces complete samples of type-I AGN. We find a completeness of 73 % for the first and 67 % for the second. Of the objects that the 2-lines mode does not classify as type-I AGN, 88 % of them have PDZ($z_{\rm spec}$) < 0.5. We check that we do not obtain PDZ($z_{\rm spec}$) > 0.5 for them including in LePHARE all the AGN templates from Salvato et al. (2009, 2011) and from the LEPHARE distribution. For the 3-lines mode we find that 60 % of the objects not confirmed as type-I AGN have PDZ ($z_{\rm spec}$) < 0.5. The rest of them are rejected because ELDAR does not detect at least 3 AGN emission lines in their photometry. It is the consequence of objects for which some of their emission lines have a EW smaller than values listed in Table 4.1, and thus the ALHAMBRA bands are not narrow enough to detect them. The EWs listed in Table 4.1 were computed for 184 quasars observed by the HST (emission lines with $\lambda_{\rm c} < 1300$ Å) and for 2000 quasars observed by the SDSS (emission lines with $\lambda_{\rm c} > 1300$ Å).

In Fig. 4.19 we display the completeness for the ALH2L and ALH3L catalogues as a function of z. For the ALH3L catalogue it grows with z and approaches the completeness for the ALH2L catalogue. It is smaller at $z < 2$ for both catalogues because the AGN continuum emission is very blue and steep in this redshift range,

and thus the emission lines require a large EW to be detected. On the other hand, the completeness is greater at $z > 2$ because the AGN continuum is less steep. We explore this issue in more detail in §4.5. We do not include the completeness at $z > 2.5$ because there are just two objects in the AGN-X sample at $z > 2.5$.

Another concern is that the type-I AGN could be best-fitted by stellar templates, and thus we would reject them. Nevertheless, no source of the AGN-X sample is best-fitted by a stellar template.

4.5 Forecasts for narrow band surveys

In this section we forecast the performance of ELDAR for surveys with narrower bands than the ALHAMBRA survey, as our method can be applied to any survey in which the bands are narrow enough for isolating emission lines from the continuum.

There are several surveys that incorporate contiguous bands narrower than the ALHAMBRA bands, such as SHARDS (25 bands of FWHM $\simeq 170$ Å), PAUS (40 bands of FWHM $\simeq 130$ Å), and J-PAS (54 bands of FWHM $\simeq 140$ Å). As the data from all of these surveys is not publicly available yet and we want to forecast the performance of ELDAR for different filter systems, we decided to forecast the completeness and redshift precision for J-PAS. We select this survey because J-PAS has the greatest number of bands, and thus for this survey we may find the largest differences with the results for the ALHAMBRA survey.

To estimate the performance of ELDAR detecting type-I AGN in J-PAS and to make a fair comparison with ALHAMBRA, we generate AGN-mock data for the ALHAMBRA and J-PAS filter systems. In order to do that, we convolve the template qso-0.2_84 shifted in redshift between $z = 1$ and $z = 5$ using a redshift step of $\Delta z = 0.02$ with both filter systems. Then, we generate 4 mock sources at each redshift imposing a magnitude of 19.5, 20.5, 21.5, and 22.5 in the detection band of ALHAMBRA and J-PAS, which are the F814W band for the first and the r band for the second. We note that these magnitudes correspond to different SNR in the bands of these surveys, given their different magnitude limits. Next, we compute the error in each band using a empirical relation for ALHAMBRA mock data and the J-PAS exposure time calculator for J-PAS mock data (J. Varela, private communication). Finally, we apply the 2- and 3-lines mode of ELDAR to both samples, where the only modification that we include in ELDAR for J-PAS data is that we change the redshift step of LePHARE from 0.01 to 0.001. This is done because J-PAS includes narrower and more numerous contiguous bands than the ALHAMBRA survey, and thus we expect a higher redshift precision for this survey (Benítez et al. 2009b).

In Fig. 4.20 we show the performance of ELDAR for detecting type-I AGN in ALHAMBRA and J-PAS as a function of the redshift and magnitude of the source. At low redshift, the gaps in the redshift distribution are caused by the blue and steep continuum emission of the qso-0.2_84 template, which makes more difficult to detect of emission lines. This is even more important for mock sources dimmer than 21 magnitudes, none of them are confirmed by ELDAR. Nonetheless, as we can see in Fig. 4.11, at $z < 1.5$ we detect plenty of ALHAMBRA sources with F814W > 21. This is because the qso-0.2_84 template has a very steep continuum, which reduces

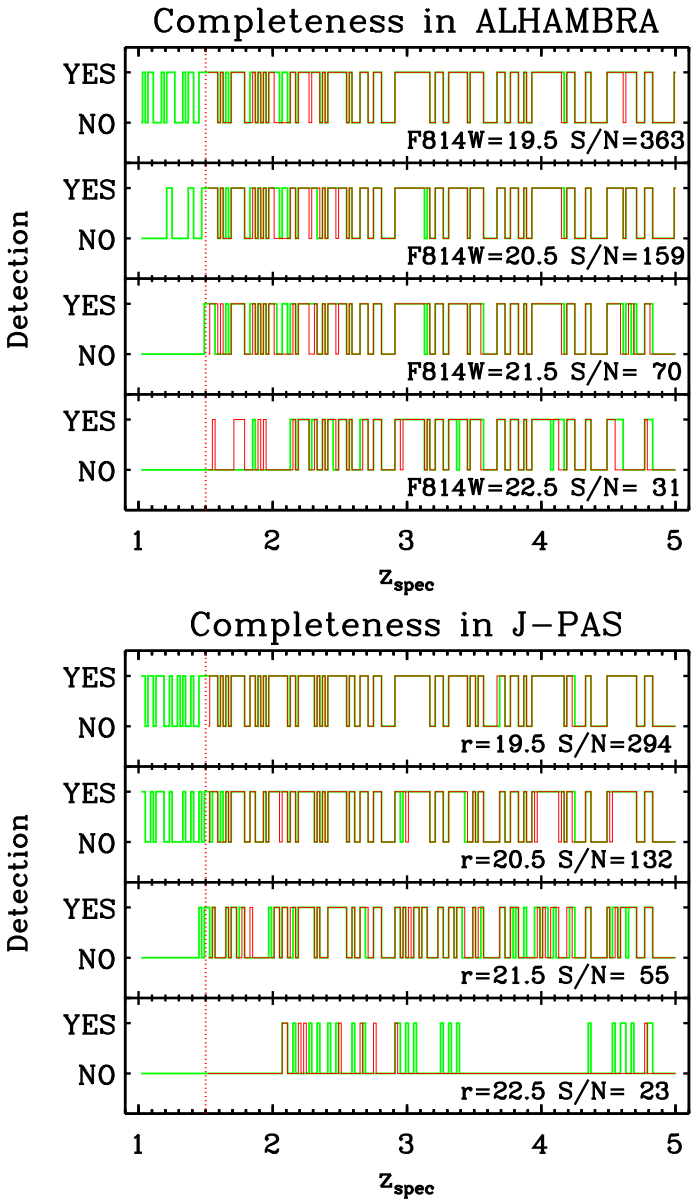


Figure 4.20: Detection of type-I AGN in ALHAMBRA and J-PAS using the 2- and 3-lines methods of ELDAR as a function of the magnitude in the detection band and the redshift of the mock source. The results are generated by convolving the synthetic template qso-0.2_84 with the ALHAMBRA and J-PAS filter systems. Green and red lines show the results for the 2-lines and 3-lines modes, respectively. There are bright objects which are not detected due to emission lines falling in between two bands. The number of detections for the faintest objects is smaller because the SNR is not large enough to detect all the AGN emission lines that ELDAR looks for.

the efficiency of our methodology.

The no-detection of bright objects at $z > 2$ at certain redshifts are due to Ly α falling in between two bands. Whereas this is an important issue for ALHAMBRA, it could be solved for J-PAS data. If we introduce a redshift-dependent continuum or we model it using the bands which are adjacent to the band where the lines fall, we could confirm these sources. On the other hand, the no-detection of dim objects is because the SNR required to detect the AGN lines gathered in Table 4.1 in their photometric data is not large enough.

Using the 3-lines mode of ELDAR, we achieve a redshift precision of $\sigma_{\rm NMAD} = 0.48\,\%$ for mock ALHAMBRA data and $\sigma_{\rm NMAD} = 0.21\,\%$ for mock J-PAS data. As we get $\sigma_{\rm NMAD} = 0.84\,\%$ for real ALHAMBRA data (see Table 4.4), we forecast a precision of $\sigma_{\rm NMAD} = 0.37\,\%$ for J-PAS, which is similar to the one expected for J-PAS galaxies (Benítez et al. 2014).

4.6 Summary and conclusions

The emergence of medium- and narrow-band photometric surveys make necessary to develop new techniques to fully exploit their data. In this work we presented ELDAR, a new method that enables the secure identification of unobscured AGN and the precise computation of their redshifts. As input our method only employs the multi-band photometric data of the sources to be analysed. Then, it takes advantage of these low-resolution spectra to detect AGN emission lines, which enables the unambiguously confirmation of sources as AGN.

In order to test the performance of ELDAR, we applied it to the publicly available data from the ALHAMBRA survey, which covered $\simeq 3\,{\rm deg}^2$ of the northern sky with 20 contiguous medium-bands of FWHM $\simeq 300\,$Å. Given the characteristics of this survey, we tuned ELDAR to detect type-I AGN. Specifically, we defined two different configurations of our method, the first requiring the detection of at least two AGN emission lines and the second of at least three. Running the both modes of ELDAR on the ALHAMBRA data, we ended up with 585 and 494 sources at $z > 1$ with F814W < 22.5 and at $z > 1.5$ with F814W < 23, respectively, where 461 and 408 of them are new. Then, to characterise these catalogues, we ran the two configurations on samples of spectroscopically-known type-I AGN and galaxies in the ALHAMBRA fields, estimating a completeness of 73 % and 67 %, a redshift precision of $\sigma_{\rm NMAD} = 0.97\,\%$ and $\sigma_{\rm NMAD} = 0.84\,\%$, and a galaxy contamination of 31 % and 9 %, respectively. Moreover, the galaxy contamination is zero at $z > 2$ for the both catalogues.

Consequently, ELDAR improves on traditional photometric approaches, e.g. colour-colour selection techniques. Moreover, ELDAR does not require additional data from X-ray, radio, nor variability studies, whereas this is needed for other works in order to detect AGN in multi-band surveys (e.g., Salvato et al. 2009; Cardamone et al. 2010; Luo et al. 2010; Salvato et al. 2011; Fotopoulou et al. 2012).

Finally, we forecast the performance of ELDAR in surveys with narrower bands. We analysed the particular case of the upcoming J-PAS survey, which consists of 54 narrow-bands of FWHM $\simeq 140\,$Å. Using J-PAS mock data, we estimated a redshift precision of $\sigma_{\rm NMAD} = 0.37\,\%$ and approximately the same completeness as ALHAMBRA,

which is thanks to its narrower bands (J-PAS is shallower than ALHAMBRA). Furthermore, it will be possible to improve on the completeness for the J-PAS data by using an optimal estimation of the AGN continuum emission for this survey.

Appendix A: AGN examples

To illustrate the objects of the AGN-S sample that ELDAR confirms and rejects and why it does so, in Figs. 4.21–4.26 we display the photometric data of various objects and we discuss whether they fulfil all ELDAR criteria or not. In the figures we use arrows to point to the bands where ELDAR detects AGN emission lines. We indicate the name of the lines according to $z_{\rm spec}$.

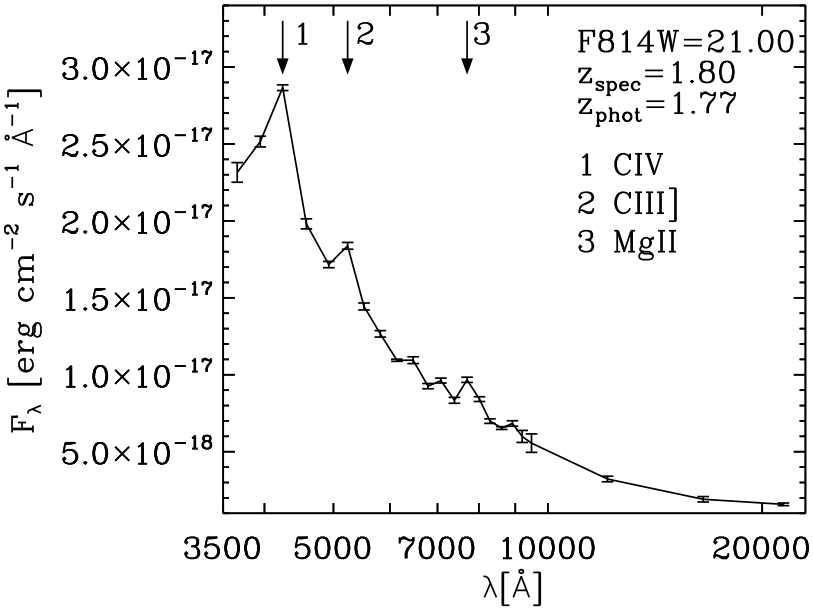


Figure 4.21: Source of the AGN-S sample at $z_{\rm spec} = 1.80$ that it is classified as type-I AGN by the 2- and 3-lines modes of ELDAR. Arrows point to the bands where AGN emission lines are confirmed.

In Fig. 4.21 we show an object of the AGN-S sample at $z_{\rm spec} = 1.80$. After applying the 2- and 3-lines modes of ELDAR, we find that it is correctly classified as AGN by both. This is because our method detects C IV in the 3rd band, C III] in the 5th band, and Mg II in the 14th band. In addition, the redshift that ELDAR assigns to this object, $z_{\rm phot} = 1.77$, is compatible with its spectroscopic redshift, $z_{\rm phot} = 1.78$.

In Fig. 4.22 we plot a source of the AGN-S sample at $z_{\rm spec} = 1.83$. This object is classified as AGN just by the 2-lines mode of ELDAR. This is because our method detects C IV in the 3rd band, Mg II in the 15th band, but not C III] because it falls between the 6th and 7th bands. The photometric redshift computed by ELDAR, $z_{\rm phot} = 1.83$, is the same as its spectroscopic redshift, $z_{\rm spec} = 1.83$.

In Fig. 4.23 we display an object of the AGN-S sample at $z_{\rm spec} = 3.52$ that is classified as type-I AGN by the 3-lines mode of ELDAR. This is because our method

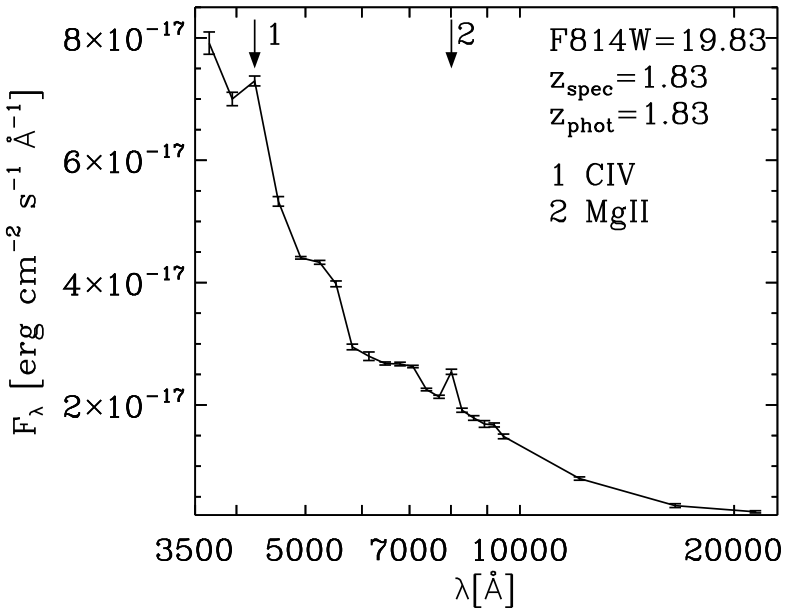


Figure 4.22: Object of the AGN-S sample at $z_{\rm spec} = 1.83$ that it is classified as type-I AGN just by the 2-lines mode of ELDAR. This is because our code does not detect the line C III] at 5404Å.

detects the complex O VI+Ly β in the 4th band, Ly α in the 8th band, and C III] in the 17th band. On the other hand, the line C IV is not detected because it falls between the 12th and 13th bands. According to $z_{\rm spec}$, the central wavelength of the lines Ly α and C IV should be very close to the central wavelength of the 7th and 12th band, respectively. However, the first falls between the 7th and 8th band and the second between the 12th and 13th band. This is because AGN emission lines may be shifted with respect to their rest-frame wavelength and/or have anisotropic profiles (see Vanden Berk et al. 2001), where these effects can modify the band where they fall. As a consequence, the photometric redshift computed for this source, $z_{\rm phot} = 3.63$, is $\simeq 3\,\%$ greater than its spectroscopic redshift.

In Fig. 4.24 we show the only object of the AGN-S sub-sample with $z_{\rm spec} > 2.75$ not confirmed as type-I AGN by the 3-lines mode. However, it is classified as type-I AGN by the 2-lines mode. This is because our code does not detect C IV, which should fall in the 10th band, nor C III], which should fall in the 15th band. It is the consequence of the ALHAMBRA bands not been narrow enough for detecting these lines. The lack of these lines causes the computed photometric redshift, $z_{\rm phot} = 3.14$, to be $\simeq 8\,\%$ greater than the spectroscopic redshift for this object, $z_{\rm spec} = 2.91$.

In the Figs. 4.22 and 4.24 we have shown a low-z and a high-z spectroscopically-known object that are not classified as type-I AGN by the 3-lines mode of ELDAR. Objects like these explain why the 3-lines mode has a smaller completeness than the 2-lines mode. In the following, we will show some examples of spectroscopically-known objects for which ELDAR produces catastrophic redshift solutions.

In Fig. 4.25 we display a source of the AGN-S sample at $z_{\rm spec} = 1.69$ classified

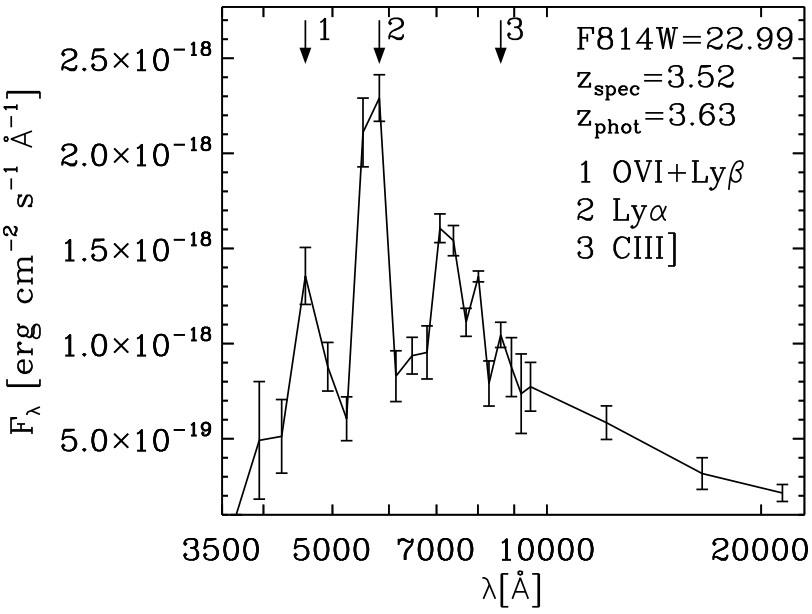


Figure 4.23: Source of the AGN-S sample at $z_{\rm spec} = 3.52$ that is classified as type-I AGN by the 3-lines mode of ELDAR.

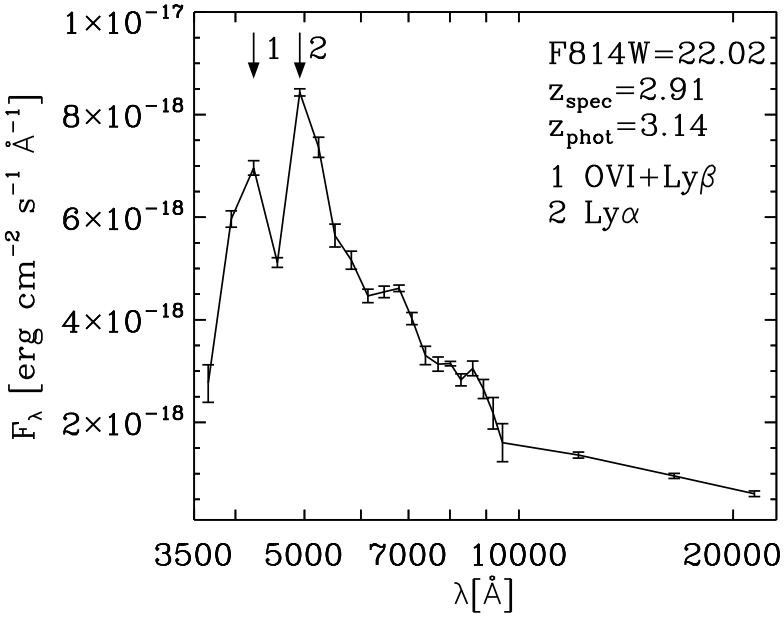


Figure 4.24: Object of the AGN-S sample at $z_{\rm spec} = 2.91$ that it is only classified as type-I AGN by the 2-lines mode. It is not confirmed by the 3-lines mode because our code does not detect emission in C IV at 6407Å nor in C III] at 7895Å.

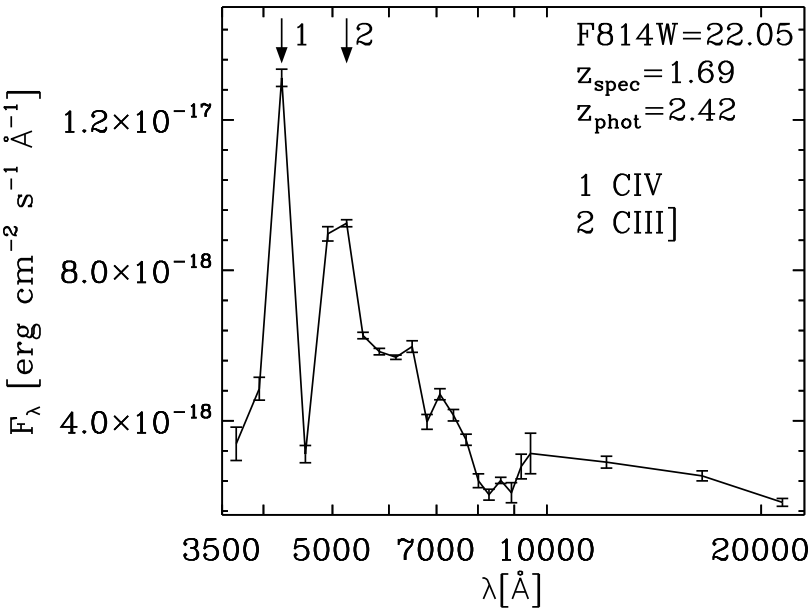


Figure 4.25: Object of the AGN-S sample at $z_{\rm spec} = 1.69$ that it is classified as type-I AGN by the 2- and 3-lines modes of ELDAR at $z_{\rm phot} = 2.42$. Our code produces a catastrophic redshift solution for this source because the position of the pair {Ly α and C IV} at $z = 2.42$ is degenerated with the position of the pair {C IV and C III]} at $z = 1.69$. In addition, ELDAR detects a spurious line in the 9th band that it is confused with C III] at $z_{\rm phot} = 2.42$.

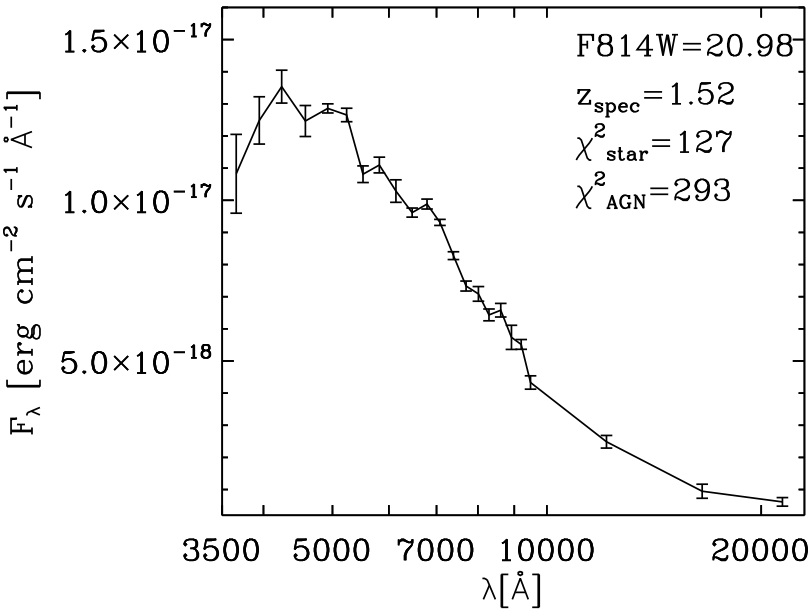


Figure 4.26: Only source of the AGN-S sample that it is best-fitted by a stellar template. We cannot see clear emission lines in the ALHAMBRA photometric data.

as type-I AGN by the 2- and 3-lines modes at $z_{\rm phot} = 2.42$. Thus, this object is an outlier according to our definition (see §4.4.2). This is because i) PDZ($z_{\rm spec}$) < 0.5 and ii) there is a degeneracy between the pair {Ly α and C IV} at $z = 2.42$ and the pair {C IV and C III]} at $z = 1.69$. This source is also confirmed by the 3-lines mode because C III] is confused with a spurious line detected in the 9th band.

In Fig. 4.26 we display the only object of the AGN-S sample best-fitted by a stellar template. This object is at $z_{\rm spec} = 1.52$ and it does not show any clear emission lines. The best-fitting AGN templates has a χ^2 more than twice the χ^2 of the best-fitting stellar templates. Even if this object is not best-fitted by an AGN template, it will not be confirmed as type-I AGN because ELDAR does not detect any AGN emission lines. No objects from the AGN-X sample are best-fitted by stellar templates.

Appendix B: Dependence of the results on the criteria adopted in ELDAR

In §4.2 and §4.3.3 we introduced multiple parameters in the configuration of ELDAR. In this section we show the dependence of the results for the AGN-S and GAL-S samples on these criteria. We will apply the 2- and 3-lines modes to these samples, modifying just one free parameter at a time. In all the tables we underline the results for the fiducial configuration of ELDAR.

Table 4.6: Results for the AGN-S and GAL-S samples as a function of the PDZ cut-off.

PDZ	Mode	Compl.(%)	$\sigma_{\rm NMAD}$(%)	η(%)	Galaxies
0.90	2-lines	72.0	0.97	7.1	4
	3-lines	64.6	0.84	8.1	1
0.50	2-lines	<u>72.8</u>	<u>0.97</u>	<u>8.1</u>	<u>4</u>
	3-lines	<u>64.6</u>	<u>0.84</u>	<u>8.1</u>	<u>1</u>
0.01	2-lines	73.5	0.98	9.0	4
	3-lines	65.6	0.84	7.9	1

Notes. Underlined numbers denote fiducial values for the 2- and 3-lines modes of ELDAR.

We introduced a PDZ cut-off of 0.5 in ELDAR to reject redshift solutions for which the χ^2 is very low. In Table 4.6 we gather the results for the AGN-S and GAL-S samples using different values of the PDZ cut-off. The quality of the ALH2L and ALH3L catalogues is largely independent of the value of this parameter. This is because most of the objects with F814W < 22.5 have only one peak with PDZ > 0.5.

Another criterion that we included in ELDAR is that the flux in the band where the Ly α line falls has to be 75 % greater than the flux in the rest of the bands. In Table 4.7 we present the results for the AGN-S and GAL-S using different percentages. We find that increasing this percentage the completeness is reduced.

Table 4.7: Results for the AGN-S and GAL-S samples as a function of the Ly α criterion.

Ly α	Mode	Compl.(%)	$\sigma_{\rm NMAD}$(%)	η(%)	Galaxies
1.25	2-lines	68.4	0.96	6.4	4
	3-lines	59.4	0.83	5.3	1
0.75	2-lines	72.8	0.97	8.1	4
	3-lines	64.6	0.84	8.1	1
0.25	2-lines	73.5	0.97	8.0	4
	3-lines	65.6	0.83	7.9	1

Table 4.8: Results for the AGN-S and GAL-S samples as a function of the line acceptance criterion.

$\sigma_{\rm line}$	Mode	Compl.(%)	$\sigma_{\rm NMAD}$(%)	η(%)	Galaxies
0.50	2-lines	77.2	0.99	7.6	13
0.75		77.2	0.99	7.6	11
1.00		75.0	0.98	7.8	7
1.25		75.7	0.99	8.7	4
1.50		72.8	0.97	8.1	4
1.75		71.3	0.96	8.2	1
0.50	3-lines	65.6	0.84	7.9	1
0.75		64.6	0.84	8.1	1
1.00		61.4	0.87	8.5	1
1.25		56.2	0.80	7.4	1
1.50		52.1	0.76	6.0	0
1.75		49.0	0.70	4.9	0

Notes. $\sigma_{\rm line}$ indicates minimum the number of σs that we require to emission lines.

Table 4.9: Results for the AGN-S and GAL-S samples as a function of $z_{\rm min}$.

$z_{\rm min}$	Mode	Compl.(%)	$\sigma_{\rm NMAD}$(%)	η(%)	Galaxies
1.0	2-lines	72.8	0.97	8.1	4
	3-lines	-	-	-	-
1.5	2-lines	76.8	0.83	5.5	3
	3-lines	64.6	0.84	8.1	1
2.0	2-lines	84.3	0.83	7.0	0
	3-lines	80.4	0.87	12.1	0
2.5	2-lines	79.2	0.97	0.0	0
	3-lines	79.2	0.97	5.2	0

Notes. The 3-lines mode is not defined at $z = 1$ because there are less than 3 AGN emission lines that ELDAR looks for within the ALHAMBRA wavelength coverage.

Table 4.10: Results for the AGN-S and GAL-S samples as a function of the magnitude limit in the detection band, F814W.

F814W	Mode	Compl.(%)	$\sigma_{\rm NMAD}$(%)	η(%)	Galaxies
21.5	2-lines	73.1	0.92	6.6	2
	3-lines	61.1	0.67	2.2	1
22.0	2-lines	71.3	0.92	6.9	3
	3-lines	60.2	0.68	2.0	1
22.5	2-lines	72.8	0.97	8.1	4
	3-lines	63.2	0.80	5.0	1
23.0	2-lines	73.1	0.99	9.9	4
	3-lines	64.6	0.84	8.1	1

We set different requirements to confirm emission lines for the 2- and 3-lines modes. In the 2-lines mode we established a stricter acceptance criterion than in the 3-lines mode to reduce the galaxy contamination. In Table 4.8 we display the results for the AGN-S and GAL-S sample using different acceptance criteria. We find that the smaller is the value of $\sigma_{\rm line}$, the higher is the completeness and the galaxy contamination. Moreover, the galaxy contamination strongly grows by reducing $\mathcal{N}$, and thus $\sigma_{\rm line}$ has to be carefully chosen depending on $\mathcal{N}$.

The condition of detecting at least 2 or 3 AGN emission lines to confirm objects sets a minimum redshift, $z_{\rm min}$, for the sources. In order to check whether the ELDAR's performance depends on the redshift of the sources, we apply the 2- and 3-lines modes to the AGN-S and GAL-S samples using different values of $z_{\rm min}$. In Table 4.9 we gather the results. We find that the completeness increases as a function of the redshift, and the galaxy contamination decreases. Moreover, the redshift precision is largely independent of $z_{\rm min}$.

Finally, we address the dependence of the results on the magnitude limit. In Table 4.10 we summarize the results for the AGN-S and GAL-S samples using the 2- and 3-lines modes. For both modes, the completeness does not depend strongly on the magnitude limit; however, the redshift precision grows for brighter objects and the galaxy contamination decreases.

Appendix C: Description of the ALH2L and ALH3L catalogues

The catalogues ALH2L and ALH3L are available as binary ASCII tables. The documentation is provided in a README file (column, bytes, format, units, label, description) and it is also shown in §4.4.

Notes on the catalogue columns:

1 The identification of the object, given by the format ALHXLYYY where the value of X is 2 for the ALH2L and 3 for the ALH3L catalogue, and YYY is the ID of the object. The IDs are ordered according to $z_{\rm phot}$.

2 - 4 The J2000 coordinates (right ascension, sign of the declination, and declination). The astrometry is from ALHAMBRA.

5 ELDAR photometric redshift.

6 It indicates whether the object is inside the ALHAMBRA mask (1) or not (0).

7 Index of the AGN template that best-fit the data.

8 - 9 Extinction law and colour excess of the extragalactic template that best-fit the data. The extinction is 0 for templates without extinction and 1 for the Calzetti et al. (2000) extinction law, 2 for the Allen (1976) extinction law, 3 for the Prevot et al. (1984) extinction law, and 4 for the Fitzpatrick (1986) extinction law.

10 - 11 PSF-magnitude and uncertainty in the F814W band.

12 The Stellarity parameter of SExtractor. In ALHAMBRA it does not provide accurate results for objects with F814W > 23.

13 - 50 PSF-magnitude and uncertainty in the ALHAMBRA medium-bands.

51 - 56 PSF-magnitude and uncertainty in the ALHAMBRA infrared broad-bands.

57 - 74 Number of the ALHAMBRA band where the AGN emission lines of Table 4.1 fall. We set this value to 99 for no detections and to 0 for lines outside the ALHAMBRA medium-band wavelength range. For detected lines we also include the decimal logarithm of the SNR in the band where they fall and decimal logarithm of the significance with which they are detected, $\rm S_{lin}$, defined as:

$$\mathrm{S_{lin}} = \min \begin{cases} \dfrac{F_{\rm cen} - F_{\rm blue}}{S_{\rm cen}} - \sigma_{\rm line}, \\ \dfrac{F_{\rm cen} - F_{\rm red}}{S_{\rm cen}} - \sigma_{\rm line}, \\ \dfrac{F_{\rm cen} - F_{\rm blue}}{S_{\rm cen}} - \sigma_{\rm line}\dfrac{S_{\rm blue}}{S_{\rm cen}}, \\ \dfrac{F_{\rm cen} - F_{\rm red}}{S_{\rm cen}} - \sigma_{\rm line}\dfrac{S_{\rm red}}{S_{\rm cen}}. \end{cases} \tag{4.7}$$

Table 4.11. Byte-by-byte description of the ALH2L and ALH3L catalogues.

Column	Bytes	Format	Units	Label	Description
1	1-8	A8	–	ID	Identification number
2	10-17	F8.4	deg	RA	Right Ascension J2000 [0, 360]
3	19	A1	–	DE-	Declination J2000 (sign)
4	20-26	F7.4	deg	DEC	Declination J2000 [-90, 90]
5	28-32	F5.3	–	Z	Photometric redshift
6	34	I1	–	MASK	Mask [0 outside, 1 inside]
7	36-37	I2	–	TEMP	Best-fit extragalactic template
8	39-42	F4.2	–	EXTB	Best-fit colour excess
9	44-49	F6.3	mag	F814W	F814W magnitude
10	51-55	F5.3	mag	eF814W	F814W uncertainty
11	57-60	F4.2	–	STELL	SExtractor Stellarity parameter [1 point-like, 0 extended]
12	62-68	F7.3	mag	F365W	F365W magnitude
13	70-76	F7.3	mag	eF365W	F365W uncertainty
14	78-84	F7.3	mag	F396W	F396W magnitude
15	86-92	F7.3	mag	eF396W	F396W uncertainty
16	94-100	F7.3	mag	F427W	F427W magnitude
17	102-108	F7.3	mag	eF427W	F427W uncertainty
18	110-116	F7.3	mag	F458W	F458W magnitude
19	118-124	F7.3	mag	eF458W	F458W uncertainty
20	126-132	F7.3	mag	F489W	F489W magnitude
21	134-140	F7.3	mag	eF489W	F489W uncertainty
22	142-148	F7.3	mag	F520W	F520W magnitude
23	150-156	F7.3	mag	eF520W	F520W uncertainty
24	158-164	F7.3	mag	F551W	F551W magnitude
25	166-172	F7.3	mag	eF551W	F551W uncertainty
26	174-180	F7.3	mag	F582W	F582W magnitude
27	182-188	F7.3	mag	eF582W	F582W uncertainty
28	190-196	F7.3	mag	F613W	F613W magnitude
29	198-204	F7.3	mag	eF613W	F613W uncertainty
30	206-212	F7.3	mag	F644W	F644W magnitude
31	214-220	F7.3	mag	eF644W	F644W uncertainty
32	222-228	F7.3	mag	F675W	F675W magnitude
33	230-236	F7.3	mag	eF675W	F675W uncertainty
34	238-244	F7.3	mag	F706W	F706W magnitude

Table E1. Continued.

Column	Bytes	Format	Units	Label	Description
35	246-252	F7.3	mag	eF706W	F706W uncertainty
36	254-260	F7.3	mag	F737W	F737W magnitude
37	262-268	F7.3	mag	eF737W	F737W uncertainty
38	270-276	F7.3	mag	F768W	F768W magnitude
39	278-284	F7.3	mag	eF768W	F768W uncertainty
40	286-292	F7.3	mag	F799W	F799W magnitude
41	294-300	F7.3	mag	eF799W	F799W uncertainty
42	302-308	F7.3	mag	F830W	F830W magnitude
43	310-316	F7.3	mag	eF830W	F830W uncertainty
44	318-324	F7.3	mag	F861W	F861W magnitude
45	326-332	F7.3	mag	eF861W	F861W uncertainty
46	334-340	F7.3	mag	F892W	F892W magnitude
47	342-348	F7.3	mag	eF892W	F892W uncertainty
48	350-356	F7.3	mag	F923W	F923W magnitude
49	358-364	F7.3	mag	eF923W	F923W uncertainty
50	366-372	F7.3	mag	F954W	F954W magnitude
51	374-380	F7.3	mag	eF954W	F954W uncertainty
52	382-388	F7.3	mag	J	J magnitude
53	390-396	F7.3	mag	eJ	J uncertainty
54	398-404	F7.3	mag	H	H magnitude
55	406-412	F7.3	mag	eH	H uncertainty
56	414-420	F7.3	mag	Ks	Ks magnitude
57	422-428	F7.3	mag	eKs	Ks uncertainty
58	430-431	I2	–	LINE1	Band where the O VI+Ly β complex is detected [2,19]
59	433-438	F6.3	–	SNLINE1	$\log_{10}(SNR)$ in the band where the O VI+Ly β complex is detected
60	440-445	F6.3	–	SLINE1	$\log_{10}(S_{lin})$ in the band where the O VI+Ly β complex is detected
61	447-448	I2	–	LINE2	Band where the Ly α line is detected [2,19]
62	450-455	F6.3	–	SNLINE2	$\log_{10}$(SNR) in the band where the Ly α line is detected
63	457-462	F6.3	–	SLINE2	$\log_{10}(S_{lin})$ in the band where the Ly α line is detected
64	464-469	I2	–	LINE3	Band where the Si IV+O IV]

Table E1. Continued.

Column	Bytes	Format	Units	Label	Description
					complex is detected [2,19]
65	471-476	F6.3	–	SNLINE3	$\log_{10}$(SNR) in the band where the Si IV+O IV] complex is detected
66	478-479	F6.3	–	SLINE3	$\log_{10}(S_{lin})$ in the band where the Si IV+O IV] complex is detected
67	481-482	I2	–	LINE4	Band where the C IV line is detected [2,19]
68	484-489	F6.3	–	SNLINE4	$\log_{10}$(SNR) in the band where the C IV line is detected
69	491-496	F6.3	–	SLINE4	$\log_{10}(S_{lin})$ in the band where the C IV line is detected
70	498-499	I2	–	LINE5	Band where the C III] line is detected [2,19]
71	501-506	F6.3	–	SNLINE5	$\log_{10}$(SNR) in the band where the C III] line is detected
72	508-513	F6.3	–	SLINE5	$\log_{10}(S_{lin})$ in the band where the C III] line is detected
73	515-516	I2	–	LINE6	Band where the Mg II line is detected [2,19]
74	518-523	F6.3	–	SNLINE6	$\log_{10}$(SNR) in the band where the Mg II line is detected
75	525-530	F6.3	–	SLINE6	$\log_{10}(S_{lin})$ in the band where the Mg II line is detected

5 Summary and conclusions

"[...] and in the eyes of the people there is the failure; and in the eyes of the hungry there is a growing wrath. In the souls of the people the grapes of wrath are filling and growing heavy, growing heavy for the vintage".

—John Steinbeck, *The Grapes of Wrath*

The ΛCDM model makes robust and detailed predictions for multiple cosmological observables from the late to the early universe, and thus it may be precisely constrained with observations. Nowadays, multiple ongoing and future galaxy surveys aim at measuring the growth and expansion history of the universe with sub-percent precision, and to unveil the nature of DM. In this thesis we addressed several challenges that these surveys will face to extract unbiased cosmological information from the galaxy clustering and WL analyses. Furthermore, we introduced a new methodology to identify bright tracers of the matter density field at high-z (AGN), which may be employed to do cosmology.

We divided the body of the thesis in three chapters according to their purpose. In what follows we summarise our main findings:

- As we noticed in Chapter 1, a precise understanding of the connection between galaxies and the DM density field is very important to extract unbiased cosmological constraints from galaxy surveys. In Chapter 2 we studied and modelled this relation using SHAM, an algorithm that bijectively associates galaxies to DM haloes. We found that:
  - All current SHAM implementations link galaxies to DM haloes employing properties of the DM haloes that are affected by numerical artefacts or undesired physical effects (see Fig. 2.1). To overcome these issues, we defined a new property, $V_{\rm relax}$. Using data from the cosmological hydrodynamical simulation EAGLE, in Fig. 2.4 we showed that $V_{\rm relax}$ is the DM halo parameter most strongly correlated with the galaxy stellar mass.

  - Taking advantage of the EAGLE data, we generated a new implementation of SHAM using $V_{\rm relax}$. This implementation does not include free parameters unlike others from the literature, and it only introduces a scatter between the galaxy and DM halo parameters given by the EAGLE data. Thus, the results of our implementation depend on the hydrodynamical simulation from which the scatter is measured.
  - Then, we populated a DM only simulation with galaxies using our new SHAM implementation. We found that in redshift-space the galaxies produced by SHAM showed the same clustering as the EAGLE galaxies to within statistical errors (see Fig. 2.8).
  - We detected the presence of galaxy assembly bias in EAGLE, which was the first time found in a hydrodynamical simulation. In EAGLE, this effect increases the amplitude of the galaxy clustering expected from simple HOD analyses by about 25 %, and we demonstrated that our SHAM implementation approximately captures its impact. In Fig. 2.9 we displayed these results.
  - Finally, using EAGLE data, in Fig. 2.11 we showed that, by switching off the star-formation in satellite galaxies, the amplitude of the 2PCF at 1 Mpc is suppressed by 30 % with respect to the fiducial model. On the other hand, we discovered that the amplitude of the 2PCF is increased by 15 % at the same scale by inhibiting the stripping of stars in satellites. Therefore, different prescriptions for small-scale physics significantly modify the resulting galaxy clustering.

- We mentioned in the introduction that galaxy surveys are divided into photometric, spectro-photometric, and spectroscopic surveys according to the strategy that they employed to scan the sky. While the first introduce large errors in the measured redshifts, the second may reach sub-percent level precision, and the third may even surpass it. This is important, because it enables the measurement of the three dimensional galaxy clustering. In Chapter 3 we precisely explored the effect of redshift errors on the galaxy clustering in Fourier space, with an emphasis on the BAO feature. The main results of our investigation were:
  - We developed a complete theoretical methodology to model the impact of redshift errors on the measured power spectrum multipoles and their variances. Then, we confronted our predictions with results from hundreds of cosmological simulations produced by us. In Figs. 3.1 and 3.2 we showed that our theoretical expressions capture the results from the simulations to within 5 %, where the main source of uncertainty comes from our modelling of the RSD. Moreover, in Fig. 3.5 we displayed the signal-to-noise ratio for the monopole. We showed that on scales $k\,\sigma_{\rm eff} \simeq 1$ it is greater for samples with sub-percent redshift errors than for samples with no errors, as long as the contribution of the shot-noise is negligible.
  - We discovered that redshift errors reduce the contribution of the power spectrum modes along the line-of-sight when conducting its angular average. As these modes are more strongly suppressed than the ones perpendicular to the line-of-sight, this translates into a better precision detecting the BAO feature

from samples with sub-percent errors if the contribution of the shot-noise is small (see Fig. 3.7).

- We found that in redshift space the information encoded in the BAO feature is scale-dependent, and that its dependence on the Hubble parameter decreases with the redshift error and the large-scale bias of the sample.
- We introduced a complete framework to extract the position of the BAO scale from galaxy surveys with redshift errors. Moreover, we theoretically computed the dependence of the uncertainty associated to measuring the BAO scale on the large-scale bias, redshift error, and number density of the galaxy sample; the underlying cosmology; and the volume of the survey. Then, in Fig. 3.12 we showed that our analytic expression precisely captures the results from our set of simulations.
- Finally, in Fig. 3.15 we displayed forecasts for the precision measuring cosmological information from future spectro-photometric galaxy surveys. Assuming that the number density grows linearly with the redshift error, we found that samples with no errors do not necessarily produce the strictest cosmological constraints. This is because the FoM does not show a monotonic behaviour with the redshift error and the number density.

- In Chapter 1 we commented that spectroscopic galaxy surveys like eBOSS and DESI are expected to directly measure the BAO scale from the clustering of high-z quasars. In Chapter 4 we introduced ELDAR, a new method to detect unobscured AGN and to compute their redshift using data from spectro-photometric surveys. As we showed, ELDAR will enable the employment of high-z tracers of the matter density field to constrain cosmology in this type of surveys. In what follows we detail the main characteristics and outcomes of our methodology:
  - ELDAR starts by preselecting the sources that we want to classify. In order to do so, it runs LePHARE on them, which is a template-fitting code that enables the rejection of stars and the production of a PDZ for every extragalactic source. Then, ELDAR securely confirms as AGN the sources for which it detects AGN emission lines in the multi-band photometry produced by spectro-photometric surveys.
  - To characterise ELDAR, we applied it to the publicly available data from the ALHAMBRA survey. We chose this survey because it employed 20 contiguous bands of FWHM $\simeq 300\,$Å to observe the sky. Given the width of the ALHAMBRA bands, we tuned ELDAR to detect type-I AGN. We then defined two different configurations of ELDAR, the first prioritising completeness and the second a low galaxy contamination. After that, we ran both configurations on the ALHAMBRA data. We ended up with 585 type-I AGN with F814W < 22.5 at $z > 1$ and 494 sources with F814W < 23 at $z > 1.5$, respectively, where 461 and 408 of them are not spectroscopically-known.
  - To estimate the completeness, redshift precision, and galaxy contamination of the previous samples, we applied the two ELDAR configurations to spectroscopically-known type-I AGN observed by ALHAMBRA. We found a completeness of 73 % and 67 %, a redshift precision of $\sigma_{\rm NMAD} = 0.97\,\%$ and $\sigma_{\rm NMAD} = 0.84\,\%$, and a galaxy contamination of 31 % and 9 %, respec-

tively. Therefore, ELDAR improves on traditional approaches and it does not require additional data from other wavelengths or variability studies to confirm the sources nor to reduce the galaxy contamination.

All our findings will contribute to a better exploitation of galaxy surveys. Whereas the results of Chapter 2 are general for all types of surveys, the outcome of Chapter 3 and 4 are especially important for spectro-photometric surveys, such as PAUS and J-PAS. In the next chapter we will discuss prospects for future work along the lines of research pursued in this thesis.

6 Future work

Your task is not to foresee the future, but to enable it.

—Antoine de Saint Exupèry, *Citadelle* or *The Wisdom of the Sands*

In this thesis we addressed some challenges that galaxy surveys face to extract unbiased cosmological information. In addition, we developed a new method to detect AGN and to compute their redshift, which open the possibility to set cosmological constraints at high-z. Here, we introduce ongoing and future projects that complement the lines of investigation pursued in the previous chapters.

6.1 Optimise SHAM for emission line galaxies

In Chapter 2 we developed a new SHAM implementation that linked galaxies to DM haloes using their stellar masses. Consequently, that implementation can be only used when it is feasible to obtain the stellar mass of the galaxies employed. Nonetheless, this is not always possible. The stellar mass is related to the intrinsic luminosity of the galaxy continuum emission. Whereas the estimation of the continuum is straightforward for bright galaxies, for faint galaxies it is very difficult and sometimes impossible. For instance, spectroscopic surveys can detect extremely faint star-forming galaxies thanks to the strength of their emission lines, but they cannot discern their continuum emission. On the other hand, these surveys can estimate other galaxy properties, e.g. the Star Formation Rate (SFR) is related to the strength of the galaxy emission lines. As a consequence, it would be useful to elaborate other SHAM implementations that employ other galaxy properties, such as the SFR.

During the summer of 2016 I visited Professor Joop Schaye in Leiden. We started a project to optimise the SHAM model introduced in Chapter 2 to connect galaxies to DM haloes according to their SFR. In order to do this, we are using data from the EAGLE suite and $V_{\rm relax}$ as DM halo property. The preliminary results are promising, we have found that there is a strong correlation between the SFR and $V_{\rm relax}$ for central

galaxies. However, this correlation is weaker for satellite galaxies, and we need to introduce a second DM halo property to strengthen this correlation. There are some obvious choices, such as the DM halo mass or the time since a satellite galaxy became a satellite. Consequently, there is still room for investigation.

The applications of this project are similar to the ones outlined in Chapter 2 for a SHAM implementation based on stellar mass. In addition, we aim at using this new implementation to estimate the properties of the DM haloes that harbour the emission line galaxies observed by a Multi Unit Spectroscopic Explorer (MUSE) proposal of Professor Schaye.

6.2 Constraining galaxy formation models using SHAM

As we mentioned in Chapter 5, one tentative application of the SHAM implementation introduced in Chapter 2 is to constrain galaxy formation models. This is possible because our implementation includes a scatter between the galaxy stellar mass and the DM halo property employed. In Chapter 2 it was measured from the largest hydrodynamical simulation of the EAGLE suite; however, it can be measured from other hydrodynamical simulations. If we measure this scatter from simulations that employ different recipes for unresolved physical processes, we will be able to tune our SHAM implementation to reproduce the same galaxy clustering as on these simulations.

To find out which prescriptions describe more precisely the physics of galaxy formation and evolution, we should populate DM-only simulations with SHAM employing different scatters. Then, we ought to compute the clustering of the galaxies that each implementations produce, where the galaxy clustering more similar to the one measured from observations will determine the hydrodynamical simulation that uses the most accurate physical recipes.

As the dispersion between the galaxy and DM halo properties may be computed to within a wide interval of redshifts, the previous procedure can be used to constrain galaxy formation models from the late to the early universe.

6.3 Reconstruction of the linear density field under the presence of redshift errors

To extract cosmological information from BAO analyses, the predictions from linear theory have to be corrected. This is because the non-linear evolution of the matter density field shifts the position of the BAO peak with respect to the linear approximation (Crocce & Scoccimarro 2008). In Chapter 3 we found that for the volume that future galaxy surveys will sample, this shift is statistically compatible between galaxy surveys with sub-percent redshift errors.

In order to correct this shift and to increase the constraining power of the BAO, there is a widely employed procedure that consists on a reconstruction of the linear density field. Although it has been applied to several spectroscopic surveys (Eisenstein et al. 2007; Schmittfull et al. 2015), it has never been used to reconstruct

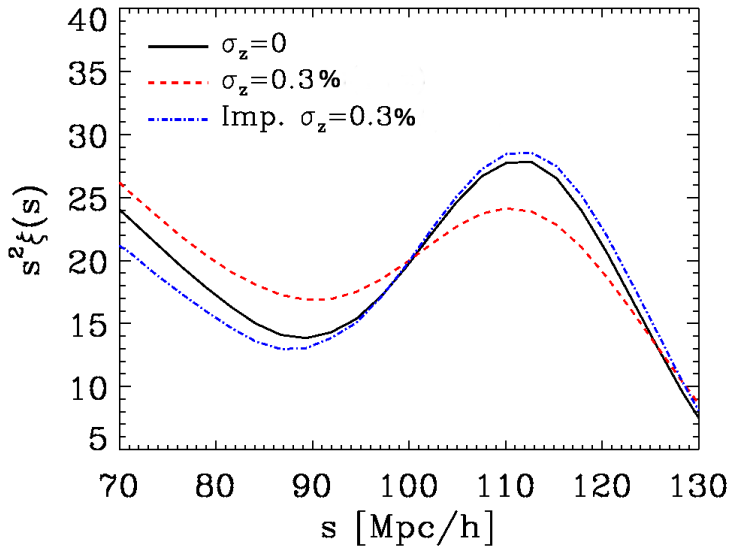


Figure 6.1: Results of a new method to reconstruct the BAO peak using galaxy samples with redshift errors. The black solid, red dashed, and blue dot-dashed lines indicate the results for galaxy samples with no errors, $\sigma_z = 0.3\,\%$, and $\sigma_z = 0.3\,\%$ after a reconstruction of the density field, respectively. As we can see, the contrast of the BAO peak for the reconstructed sample is even greater than for the sample with no errors.

the three dimensional density field of galaxy samples with redshift errors. In this project we precisely aim at developing a new reconstruction procedure to be applied on surveys that measure galaxy redshifts with noisy estimators. Our framework combines a first step in which we undo the effect of redshift errors, and a second where we apply the traditional reconstruction of the density field. We will focus on the first, as the second has been extensively explored in the literature (e.g., White 2015).

On small scales, the main effect of redshift errors is to smooth the galaxy distribution along the line-of-sight. In order to reverse this, we employ the density field as a prior for the size and direction of the corrections. We move the galaxies towards overdensities, where the displacement is drawn from a Gaussian distribution centred on the galaxy and with width equal to that of the redshift error. We have discovered that this procedure is able to correct the effect of redshift errors on the galaxy positions. In addition, we have checked that this improvement linearly grows with the logarithm of the number density of galaxies and decreases with the magnitude of the redshift errors.

In Fig. 6.1 we show the result of correcting redshift errors using a galaxy sample with $\sigma_z = 0.3\,\%$ and $n = 10^{-2}h^3\,\mathrm{Mpc}^{-3}$. Our procedure improves a $8\,\%$ the galaxy positions along the line-of-sight, which is translated into a sharper BAO peak. As we can see, this procedure approximately recovers the sharpness of the BAO peak in galaxy samples with no redshift errors.

The final step, which is still missing, is to apply the second step and check whether

both phases completely remove the shift in the BAO feature.

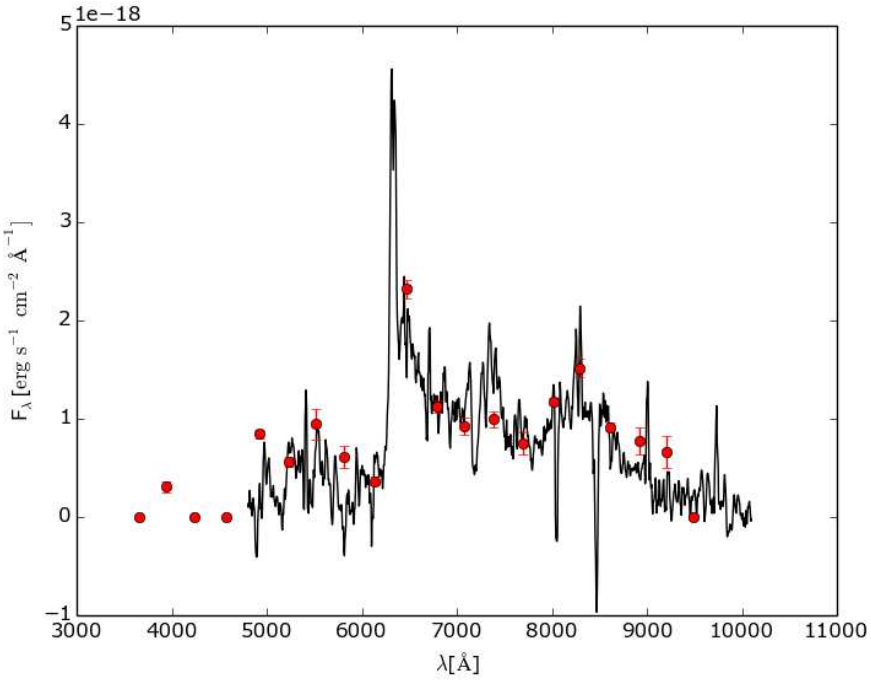


Figure 6.2: SED of a type-I AGN detected by ELDAR at $z = 4.25$ in the ALHAMBRA survey. The red dots indicate the ALHAMBRA multi-band photometry , and the black line a spectrum taken by the Gran Telescopio Canarias (GTC). This spectrum is not properly reduced, but it clearly confirms that this object is an AGN.

6.4 Spectroscopic confirmation of AGN detected at $z > 4$ by ELDAR

In order to estimate the success rate of ELDAR at high-z, Silvia Bonoli, Alexandro Ederoclite, and I applied for ten hours at the GTC to obtain the spectra of the eight type-I AGN at $z > 4$ detected by ELDAR in ALHAMBRA. In Fig. 6.2 we display the spectra of the first source observed by the GTC, which according to ELDAR is a type-I AGN at $z = 4.25$. Although this spectrum is still to be properly reduced, it is clear that this object is a type-I AGN due to the presence of typical AGN emission lines.

We have reduced the spectrum of a second object, which also looks like a type-I AGN. If all of them were successfully confirmed, it would mean that ELDAR is not only highly complete to within the redshift interval $1.5 < z < 3$ as we showed in Chapter 4, but also at higher redshifts. This is very important because the number of type-I AGN confirmed at $z > 4$ is small.

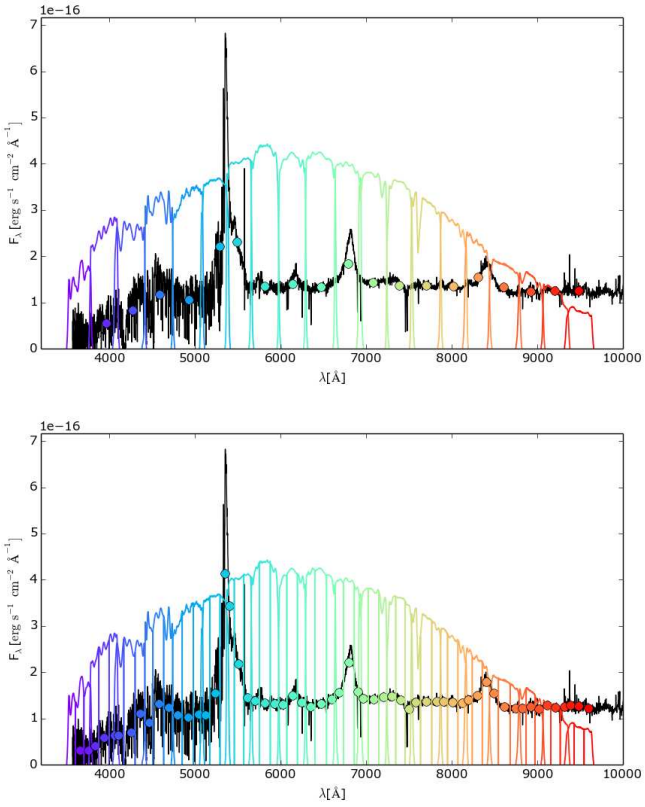


Figure 6.3: SED of a quasar observed by SDSS and convolved with the ALHAMBRA filter system (top panel) and with a narrow-band filter system with three times more filters than ALHAMBRA (bottom panel). As we can see, in narrow-band surveys the minimum width of an AGN emission line that can be detected in the multi-band photometry is smaller than in medium-band surveys.

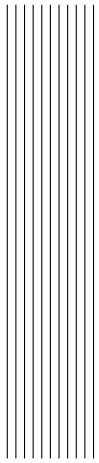
6.5 Applying ELDAR to narrow-band surveys

In Chapter 4 we tuned ELDAR to detect type-I AGN in ALHAMBRA. We note that although we only targeted type-I AGN due to the width of the ALHAMBRA bands, ELDAR is perfectly able to detect type-II AGN in surveys with narrower bands. Therefore, the obvious next step is to apply this code to other spectro-photometric surveys such as SHARDS, PAUS, and J-PAS. Whereas the core of ELDAR is largely independent of the width of the survey bands, in §4.5 we commented that the estimation of the AGN continuum may be optimised for narrow-band surveys.

In Fig. 6.3 we display the SED of a bright quasar at $z \simeq 3.2$ observed by SDSS. In the top panel we show its SED convolved with the ALHAMBRA filter system, and in the bottom panel with the filter system of a narrow-band survey like J-PAS. We can appreciate that using the filter system of the narrow-band survey, it is possible to detect narrower lines than with ALHAMBRA. In addition, it could even be feasible to estimate the width of the broadest AGN emission lines.

The efficiency of ELDAR opens the remarkable possibility of using AGN detected in narrow-band surveys to constrain cosmological parameter at high-z. We plan to pursue this in the future with the arrival new, larger, and deeper datasets, e.g. SHARDS, PAUS, and

J-PAS.



References

Abramo L. R. et al., 2012, MNRAS, 423, pp. 3251–3267.

Alam S. et al., 2016, Arxiv e-prints, 1607.03155.

Albrecht A., Steinhardt P. J., 1982, Physical review letters, 48, pp. 1220–1223.

Allen C. W., 1976, Astrophysical Quantities, London: Athlone (3rd edition).

Alpher R. A., 1948, Physical review, 74, pp. 1577–1589.

Alpher R. A., Herman R., 1948, Nature, 162, pp. 774–775.

Anderson L. et al., 2012, MNRAS, 427, pp. 3435–3467.

Angulo R. E., Baugh C. M., Frenk C. S., Lacey C. G., 2008, MNRAS, 383, pp. 755–776.

Angulo R. E., Hilbert S., 2015, MNRAS, 448, pp. 364–375.

Angulo R. E., Lacey C. G., Baugh C. M., Frenk C. S., 2009, MNRAS, 399, pp. 983–995.

Angulo R. E., Springel V., White S. D. M., Jenkins A., Baugh C. M., Frenk C. S., 2012, MNRAS, 426, pp. 2046–2062.

Angulo R. E., White S. D. M., 2010, MNRAS, 405, pp. 143–154.

Angulo R. E., White S. D. M., Springel V., Henriques B., 2014, MNRAS, 442, pp. 2131–2144.

Antonucci R., 1993, ARA&A, 31, pp. 473–521.

Aparicio Villegas T., et al., 2010, AJ, 139, pp. 1242–1253.

Arnalte-Mur P., et al., 2014, MNRAS, 441, pp. 1783–1801.

Arnouts S., Cristiani S., Moscardini L., Matarrese S., Lucchin F., Fontana A., Giallongo E., 1999, MNRAS, 310, pp. 540–556.

Babcock H. W., 1939, Lick observatory bulletin, 19, pp. 41–51.

Bacon D. J., Refregier A. R., Ellis R. S., 2000, 318, pp. 625–640.

Bahé Y. M., McCarthy I. G., 2015, MNRAS, 447, pp. 969–992.

Bahé Y. M., McCarthy I. G., Balogh M. L., Font A. S., 2013, MNRAS, 430, pp. 3017–3031.

Balbi A. et al., 2000, ApJ, 545, pp. L1–L4.

Ballinger W. E., Peacock J. A., Heavens A. F., 1996, MNRAS, 282, p. 877.

Bardeen J. M., Bond J. R., Kaiser N., Szalay A. S., 1986, ApJ, 304, pp. 15–61.

Bardeen J. M., Steinhardt P. J., Turner M. S., 1983, Phys. Rev. D, 28, pp. 679–693.

Barger A. J. et al., 2003, AJ, 126, pp. 632–665.

Behroozi P. S., Conroy C., Wechsler R. H., 2010, ApJ, 717, pp. 379–403.

Behroozi P. S., Wechsler R. H., Lu Y., Hahn O., Busha M. T., Klypin A., Primack J. R., 2014, ApJ, 787, p. 156.

Benítez N. et al., 2014, Arxiv e-prints, 1403.5237.

Benítez N. et al., 2009a, ApJ, 691, pp. 241–260.
Benítez N. et al., 2009b, ApJ, 692, pp. L5–L8.
Bertin E., Arnouts S., 1996, A&AS, 117, pp. 393–404.
Beutler F. et al., 2017a, MNRAS, 464, pp. 3409–3430.
Beutler F. et al., 2017b, MNRAS, 466, pp. 2242–2260.
Bixler J. V., Bowyer S., Laget M., 1991, A&A, 250, pp. 370–388.
Blake C., Bridle S., 2005, MNRAS, 363, pp. 1329–1348.
Blake C., Glazebrook K., 2003, ApJ, 594, pp. 665–673.
Blumenthal G. R., Faber S. M., Primack J. R., Rees M. J., 1984, Nature, 311, pp. 517–525.
Bohlin R. C., Colina L., Finley D. S., 1995, AJ, 110, p. 1316.
Bolzonella M., Miralles J.-M., Pelló R., 2000, A&A, 363, pp. 476–492.
Bond J. R., Efstathiou G., 1984, ApJ, 285, pp. L45–L48.
Brandt W. N., Hasinger G., 2005, ARA&A, 43, pp. 827–859.
Brusa M. et al., 2003, A&A, 409, pp. 65–78.
Busca N. G., et al., 2013, A&A, 552, p. A96.
Cai Y.-C., Angulo R. E., Baugh C. M., Cole S., Frenk C. S., Jenkins A., 2009, MNRAS, 395, pp. 1185–1203.
Calzetti D., Armus L., Bohlin R. C., Kinney A. L., Koornneef J., Storchi-Bergmann T., 2000, ApJ, 533, pp. 682–695.
Cardamone C. N. et al., 2010, ApJS, 189, pp. 270–285.
Chabrier G., Baraffe I., Allard F., Hauschildt P., 2000, ApJ, 542, pp. 464–472.
Civano F., Marchesi S., Comastri A., et al., 2016, ApJ, 819, p. 62.
Cole S. et al., 2005, MNRAS, 362, pp. 505–534.
Colombi S., Jaffe A., Novikov D., Pichon C., 2009, MNRAS, 393, pp. 511–526.
Conroy C., Wechsler R. H., Kravtsov A. V., 2006, ApJ, 647, pp. 201–214.
Contreras S., Baugh C. M., Norberg P., Padilla N., 2015, MNRAS, 452, pp. 1861–1876.
Crain R. A. et al., 2015, MNRAS, 450, pp. 1937–1961.
Crocce M., Scoccimarro R., 2008, Phys. Rev. D, 77, p. 023533.
Croton D. J., Gao L., White S. D. M., 2007, MNRAS, 374, pp. 1303–1309.
Dalal N., White M., Bond J. R., Shirokov A., 2008, ApJ, 687, pp. 12–21.
Dalla Vecchia C., Schaye J., 2012, MNRAS, 426, pp. 140–158.
Dalton G. et al., 2014, in Proc. SPIE, Vol. 9147, Ground-based and Airborne Instrumentation for Astronomy V, p. 91470L.
Davis M., Efstathiou G., Frenk C. S., White S. D. M., 1985, ApJ, 292, pp. 371–394.
Dawson K. S. et al., 2016, AJ, 151, p. 44.
de Jong R., 2011, The messenger, 145, pp. 14–16.
DESI Collaboration et al., 2016, Arxiv e-prints, 1611.00036.
Dolag K., Borgani S., Murante G., Springel V., 2009, MNRAS, 399, pp. 497–514.
Dolney D., Jain B., Takada M., 2006, MNRAS, 366, pp. 884–898.
Doré O. et al., 2014, Arxiv e-prints, 1412.4872.
Doré O. et al., 2016, Arxiv e-prints, 1606.07039.
Durier F., Dalla Vecchia C., 2012, MNRAS, 419, pp. 465–478.
Eifler T., Krause E., Dodelson S., Zentner A. R., Hearin A. P., Gnedin N. Y., 2015, MNRAS, 454, pp. 2451–2471.
Einstein A., 1916, Annalen der physik, 354, pp. 769–822.
Einstein A., 1917, Sitzungsberichte der königlich preußischen akademie der wissenschaften (berlin), seite 142-152.
Einstein A., de Sitter W., 1932, Proceedings of the national academy of science, 18, pp. 213–214.

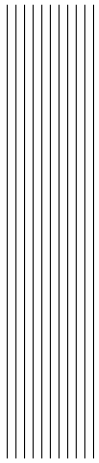
Eisenstein D. J., Seo H.-J., Sirko E., Spergel D. N., 2007, ApJ, 664, pp. 675–679.
Eisenstein D. J. et al., 2005, ApJ, 633, pp. 560–574.
Fitzpatrick E. L., 1986, AJ, 92, pp. 1068–1073.
Flesch E. W., 2015, PASA, 32, p. e010.
Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, pp. 306–312.
Fotopoulou S. et al., 2012, ApJS, 198, p. 1.
Friedmann A., 1922, Zeitschrift fur physik, 10, pp. 377–386.
Gallerani S. et al., 2010, A&A, 523, p. A85.
Gamow G., 1948, Physical review, 74, pp. 505–506.
Gao L., Springel V., White S. D. M., 2005, MNRAS, 363, pp. L66–L70.
Gao L., White S. D. M., 2007, MNRAS, 377, pp. L5–L9.
Gavignaud I., et al., 2006, A&A, 457, pp. 79–90.
Gebhardt K. et al., 2000, ApJ, 539, pp. L13–L16.
Geiss J., Reeves H., 1972, A&A, 18, p. 126.
Glazebrook K., Blake C., 2005, ApJ, 631, pp. 1–20.
Gnedin O. Y., 2003, ApJ, 582, pp. 141–161.
Gnedin O. Y., Zhao H., 2002, MNRAS, 333, pp. 299–306.
Gott, III J. R., Gunn J. E., Schramm D. N., Tinsley B. M., 1974, ApJ, 194, pp. 543–553.
Gunn J. E., 1967, 150, p. 737.
Gunn J. E., Gott, III J. R., 1972, ApJ, 176, p. 1.
Guo Q., White S., Li C., Boylan-Kolchin M., 2010, MNRAS, 404, pp. 1111–1120.
Guth A. H., 1981, Phys. Rev. D, 23, pp. 347–356.
Guth A. H., Pi S.-Y., 1982, Physical review letters, 49, pp. 1110–1113.
Hartlap J., Simon P., Schneider P., 2007, A&A, 464, pp. 399–404.
Hawking S. W., 1982, Physics letters b, 115, pp. 295–297.
Hayashi E., Navarro J. F., Taylor J. E., Stadel J., Quinn T., 2003, ApJ, 584, pp. 541–558.
Hearin A. P., Watson D. F., van den Bosch F. C., 2015, MNRAS, 452, pp. 1958–1969.
Heckman T. M., Best P. N., 2014, ARA&A, 52, pp. 589–660.
Hoaglin D. C., Mosteller F., Tukey J. W., 1983, Understanding robust and exploratory data anlysis.
Hockney R. W., Eastwood J. W., 1981, Computer Simulation Using Particles.
Hopkins P. F., 2013, MNRAS, 428, pp. 2840–2856.
Howlett C., Manera M., Percival W. J., 2015, Astronomy and computing, 12, pp. 109–126.
Hoyle F., Tayler R. J., 1964, Nature, 203, pp. 1108–1110.
Hsu L.-T., et al., 2014, ApJ, 796, p. 60.
Hu W., Haiman Z., 2003, Phys. Rev. D, 68, p. 063004.
Hubble E., 1929, Proceedings of the national academy of science, 15, pp. 168–173.
Hubble E. P., 1926, ApJ, 64.
Ilbert O. et al., 2009, ApJ, 690, pp. 1236–1249.
Jahnke K., Macciò A. V., 2011, ApJ, 734, p. 92.
Jiang L., Helly J. C., Cole S., Frenk C. S., 2014, MNRAS, 440, pp. 2115–2135.
Kaiser N., 1984, ApJ, 284, pp. L9–L12.
Kaiser N., 1987, MNRAS, 227, pp. 1–21.
Kaiser N., Wilson G., Luppino G. A., 2000.
Koda J., Blake C., Beutler F., Kazin E., Marin F., 2016, MNRAS, 459, pp. 2118–2129.
Kormendy J., Richstone D., 1995, ARA&A, 33, p. 581.
Kravtsov A. V., Gnedin O. Y., Klypin A. A., 2004, ApJ, 609, pp. 482–497.
Kuhlen M., Vogelsberger M., Angulo R., 2012, Physics of the dark universe, 1, pp. 50–93.
Lacerna I., Padilla N., 2011, MNRAS, 412, pp. 1283–1294.

Lacerna I., Padilla N., 2012, MNRAS, 426, pp. L26–L30.
Lacerna I., Padilla N., Stasyszyn F., 2014, MNRAS, 443, pp. 3107–3117.
Lacy M., et al., 2004, ApJS, 154, pp. 166–169.
Laureijs R. et al., 2011, Arxiv e-prints, 1110.3193.
Lehmer B. D. et al., 2012, ApJ, 752, p. 46.
Lemaître G., 1927, Annales de la société scientifique de bruxelles, 47, pp. 49–59.
Lemaître G., 1931, Nature, 127, p. 706.
Li Y., Mo H. J., Gao L., 2008, MNRAS, 389, pp. 1419–1426.
Lifshitz E. M., 1946, J. phys. (ussr), 10, pp. 116–129.
Lilly S. J., et al., 2009, ApJS, 184, pp. 218–229.
Linde A. D., 1982, Physics letters b, 108, pp. 389–393.
Linder E. V., 2003, Phys. Rev. D, 68, p. 083504.
Ludlow A. D., Navarro J. F., Li M., Angulo R. E., Boylan-Kolchin M., Bett P. E., 2012, MNRAS, 427, pp. 1322–1328.
Luo B. et al., 2010, ApJS, 187, pp. 560–580.
Marchesi S. et al., 2016, ApJ, 830, p. 100.
Martí P., Miquel R., Castander F. J., Gaztañaga E., Eriksen M., Sánchez C., 2014, MNRAS, 442, pp. 92–109.
Matthews T. A., Sandage A. R., 1963, ApJ, 138, p. 30.
Matute I., et al., 2012, A&A, 542, p. A20.
Matute I., et al., 2013, A&A, 557, p. A78.
Mehta K. T., Seo H.-J., Eckel J., Eisenstein D. J., Metchnik M., Pinto P., Xu X., 2011, ApJ, 734, p. 94.
Miralda-Escude J., 1991, 380, pp. 1–8.
Mo H. J., White S. D. M., 1996, MNRAS, 282, pp. 347–361.
Moles M., et al., 2008, AJ, 136, pp. 1325–1339.
Molino A. et al., 2014, MNRAS, 441, pp. 2891–2922.
Mortlock D. J. et al., 2011, Nature, 474, pp. 616–619.
Moster B. P., Somerville R. S., Maulbetsch C., van den Bosch F. C., Macciò A. V., Naab T., Oser L., 2010, ApJ, 710, pp. 903–923.
Myers S. T., Baker J. E., Readhead A. C. S., Leitch E. M., Herbig T., 1997, 485, pp. 1–21.
Nagai D., Kravtsov A. V., 2005, ApJ, 618, pp. 557–568.
Navarro J. F., Eke V. R., Frenk C. S., 1996, MNRAS, 283, pp. L72–L78.
Neto A. F. et al., 2007, MNRAS, 381, pp. 1450–1462.
Nuza S. E. et al., 2013, MNRAS, 432, pp. 743–760.
Oman K. A. et al., 2015, MNRAS, 452, pp. 3650–3665.
Padmanabhan N., White M., 2009, Phys. Rev. D, 80, p. 063508.
Pâris I. et al., 2014, Vizier online data catalog, 7270.
Pâris I. et al., 2017, A&A, 597, p. A79.
Peacock J. A., Dodds S. J., 1994, MNRAS, 267, p. 1020.
Peacock J. A., Smith R. E., 2000, MNRAS, 318, pp. 1144–1156.
Peebles P. J. E., 1982, ApJ, 263, pp. L1–L5.
Peebles P. J. E., 2001, in Astronomical Society of the Pacific Conference Series, Vol. 252, Historical Development of Modern Cosmology, Martínez V. J., Trimble V., Pons-Bordería M. J., eds., p. 201.
Peebles P. J. E., 2017, Nature astronomy, 1, p. 0057.
Peebles P. J. E., Yu J. T., 1970, ApJ, 162, p. 815.
Peng C. Y., 2007, ApJ, 671, pp. 1098–1107.
Penzias A. A., Wilson R. W., 1965, ApJ, 142, pp. 419–421.

Percival W. J. et al., 2001, MNRAS, 327, pp. 1297–1306.
Percival W. J. et al., 2007, ApJ, 657, pp. 51–55.
Percival W. J. et al., 2010, MNRAS, 401, pp. 2148–2168.
Percival W. J., White M., 2009, 393, pp. 297–308.
Perlmutter S. et al., 1999, ApJ, 517, pp. 565–586.
Peth M. A., Ross N. P., Schneider D. P., 2011, AJ, 141, p. 105.
Peñarrubia J., McConnachie A. W., Navarro J. F., 2008, ApJ, 672, pp. 904–913.
Pickles A. J., 1998, PASP, 110, pp. 863–878.
Planck Collaboration et al., 2014a, A&A, 571, p. A1.
Planck Collaboration et al., 2014b, A&A, 571, p. A16.
Planck Collaboration et al., 2016a, A&A, 594, p. A13.
Planck Collaboration et al., 2016b, A&A, 594, p. A11.
Prada F., Scóccola C. G., Chuang C.-H., Yepes G., Klypin A. A., Kitaura F.-S., Gottlöber S., Zhao C., 2016, MNRAS, 458, pp. 613–623.
Press W. H., Schechter P., 1974, ApJ, 187, pp. 425–438.
Prevot M. L., Lequeux J., Prevot L., Maurice E., Rocca-Volmerange B., 1984, A&A, 132, pp. 389–392.
Pérez-González P. G., Cava A., 2013, in Revista Mexicana de Astronomia y Astrofisica, vol. 27, Vol. 42, Revista Mexicana de Astronomia y Astrofisica Conference Series, pp. 55–57.
Raccanelli A., Bertacca D., Jeong D., Neyrinck M. C., Szalay A. S., 2016, Arxiv e-prints, 1602.03186.
Raccanelli A. et al., 2013, 436, pp. 89–100.
Read J. I., Gilmore G., 2005, MNRAS, 356, pp. 107–124.
Reddick R. M., Wechsler R. H., Tinker J. L., Behroozi P. S., 2013, ApJ, 771, p. 30.
Riess A. G. et al., 1998, AJ, 116, pp. 1009–1038.
Risaliti G., Lusso E., 2016, Arxiv e-prints, 1612.02838.
Rodney S. A. et al., 2015, 150, p. 156.
Rodríguez-Torres S. A. et al., 2016, MNRAS, 460, pp. 1173–1187.
Rosas-Guevara Y. M. et al., 2015, MNRAS, 454, pp. 1038–1057.
Ross A. J., Percival W. J., Manera M., 2015, MNRAS, 451, pp. 1331–1340.
Rubin V. C., Ford, Jr. W. K., 1970, ApJ, 159, p. 379.
Sachs R. K., Wolfe A. M., 1967, ApJ, 147, p. 73.
Salvato M., et al., 2009, ApJ, 690, pp. 1250–1263.
Salvato M., et al., 2011, ApJ, 742, p. 61.
Schaller M. et al., 2015, MNRAS, 451, pp. 1247–1267.
Schaye J. et al., 2015, MNRAS, 446, pp. 521–554.
Schaye J., Dalla Vecchia C., 2008, MNRAS, 383, pp. 1210–1222.
Schmidt B. P. et al., 1998, ApJ, 507, pp. 46–63.
Schmidt K. B., Marshall P. J., Rix H.-W., Jester S., Hennawi J. F., Dobler G., 2010, ApJ, 714, pp. 1194–1208.
Schmittfull M., Feng Y., Beutler F., Sherwin B., Chu M. Y., 2015, Phys. Rev. D, 92, p. 123522.
Scoccimarro R., Sheth R. K., Hui L., Jain B., 2001, ApJ, 546, pp. 20–34.
Sefusatti E., Crocce M., Scoccimarro R., Couchman H. M. P., 2016, MNRAS, 460, pp. 3624–3636.
Seljak U., 2000, MNRAS, 318, pp. 203–213.
Seo H.-J., Eisenstein D. J., 2003, ApJ, 598, pp. 720–740.
Seo H.-J., Eisenstein D. J., 2007, ApJ, 665, pp. 14–24.

Sereno M., Veropalumbo A., Marulli F., Covone G., Moscardini L., Cimatti A., 2015, MNRAS, 449, pp. 4147–4161.
Shankar F., Lapi A., Salucci P., De Zotti G., Danese L., 2006, ApJ, 643, pp. 14–25.
Silk J., 1968, ApJ, 151, p. 459.
Simha V., Cole S., 2013, MNRAS, 436, pp. 1142–1151.
Simha V., Weinberg D. H., Davé R., Fardal M., Katz N., Oppenheimer B. D., 2012, MNRAS, 423, pp. 3458–3473.
Smirnov Y. N., 1964, AZh, 41, p. 1084.
Smith R. E., Scoccimarro R., Sheth R. K., 2008, Phys. Rev. D, 77, p. 043525.
Springel V., 2005, MNRAS, 364, pp. 1105–1134.
Springel V. et al., 2005, Nature, 435, pp. 629–636.
Springel V., White S. D. M., Tormen G., Kauffmann G., 2001, MNRAS, 328, pp. 726–750.
Starobinsky A. A., 1982, Physics letters b, 117, pp. 175–178.
Stern D. et al., 2012, ApJ, 753, p. 30.
Sunyaev R. A., Zeldovich Y. B., 1970, Ap&SS, 7, pp. 3–19.
Suzuki N. et al., 2012, ApJ, 746, p. 85.
Sánchez A. G., Baugh C. M., Angulo R. E., 2008, MNRAS, 390, pp. 1470–1490.
Tassev S., Zaldarriaga M., Eisenstein D. J., 2013, JCAP, 6, p. 036.
Tegmark M., 1997, Physical review letters, 79, pp. 3806–3809.
Telfer R. C., Zheng W., Kriss G. A., Davidsen A. F., 2002, ApJ, 565, pp. 773–785.
Trujillo-Gomez S., Klypin A., Primack J., Romanowsky A. J., 2011, ApJ, 742, p. 16.
Tseliakhovich D., Hirata C., 2010, 82, p. 083520.
Urry C. M., Padovani P., 1995, PASP, 107, p. 803.
Uson J. M., Wilkinson D. T., 1982, Physical review letters, 49, pp. 1463–1465.
Vale A., Ostriker J. P., 2004, MNRAS, 353, pp. 189–200.
van Daalen M. P., Schaye J., McCarthy I. G., Booth C. M., Dalla Vecchia C., 2014, MNRAS, 440, pp. 2997–3010.
van de Hulst H. C., Raimond E., van Woerden H., 1957, Bull. Astron. Inst. Netherlands, 14, p. 1.
van der Hucht K. A., 2001, NAR, 45, pp. 135–232.
Van Waerbeke L. et al., 2000, 358, pp. 30–44.
Vanden Berk D. E., et al., 2001, AJ, 122, pp. 549–564.
Velliscig M., van Daalen M. P., Schaye J., McCarthy I. G., Cacciato M., Le Brun A. M. C., Dalla Vecchia C., 2014, MNRAS, 442, pp. 2641–2658.
Vogelsberger M. et al., 2014, MNRAS, 444, pp. 1518–1547.
Wang J.-M. et al., 2014, ApJ, 793, p. 108.
Watson D., Denney K. D., Vestergaard M., Davis T. M., 2011, ApJ, 740, p. L49.
Watson D. F., Berlind A. A., Zentner A. R., 2012, ApJ, 754, p. 90.
Wechsler R. H., Zentner A. R., Bullock J. S., Kravtsov A. V., Allgood B., 2006, ApJ, 652, pp. 71–84.
Weinberg D. H., Colombi S., Davé R., Katz N., 2008, ApJ, 678, pp. 6–21.
Weinberg D. H., Mortonson M. J., Eisenstein D. J., Hirata C., Riess A. G., Rozo E., 2013, Phys. Rep., 530, pp. 87–255.
Wetzel A. R., Cohn J. D., White M., 2009, MNRAS, 395, pp. 1376–1390.
Wetzel A. R., Tinker J. L., Conroy C., Bosch F. C. v. d., 2014, MNRAS, 439, pp. 2687–2700.
Wetzel A. R., Tinker J. L., Conroy C., van den Bosch F. C., 2013, MNRAS, 432, pp. 336–358.
Wetzel A. R., White M., 2010, MNRAS, 403, pp. 1072–1088.
White M., 2015, 450, pp. 3822–3828.

White M., Reid B., Chuang C.-H., Tinker J. L., McBride C. K., Prada F., Samushia L., 2015, MNRAS, 447, pp. 234–245.
White R. L. et al., 2000, ApJS, 126, pp. 133–207.
White S. D. M., Navarro J. F., Evrard A. E., Frenk C. S., 1993, Nature, 366, pp. 429–433.
White S. D. M., Rees M. J., 1978, MNRAS, 183, pp. 341–358.
Wiersma R. P. C., Schaye J., Smith B. D., 2009a, MNRAS, 393, pp. 99–107.
Wiersma R. P. C., Schaye J., Theuns T., Dalla Vecchia C., Tornatore L., 2009b, MNRAS, 399, pp. 574–600.
Wittman D. M., Tyson J. A., Kirkman D., Dell'Antonio I., Bernstein G., 2000, 405, pp. 143–148.
Wolf C., Hildebrandt H., Taylor E. N., Meisenheimer K., 2008, A&A, 492, pp. 933–936.
Wolf C. et al., 2004, A&A, 421, pp. 913–936.
Zehavi I. et al., 2005, ApJ, 630, pp. 1–27.
Zel'dovich Y. B., 1964, Soviet physics uspekhi, 6, pp. 475–494.
Zel'dovich Y. B., 1970, A&A, 5, pp. 84–89.
Zentner A. R., Hearin A. P., van den Bosch F. C., 2014, MNRAS, 443, pp. 3044–3067.
Zhao G.-B. et al., 2016, MNRAS, 457, pp. 2377–2390.
Zhu G., Zheng Z., Lin W. P., Jing Y. P., Kang X., Gao L., 2006, ApJ, 639, pp. L5–L8.
Zu Y., Zheng Z., Zhu G., Jing Y. P., 2008, ApJ, 686, pp. 41–52.
Zwicky F., 1933, Helvetica physica acta, 6, pp. 110–127.



Acronym list

ΛCDM	Lambda Cold Dark Matter xix, xxiii, 6–9, 98, 145
2PCF	Two Point Correlation Function 35, 36, 38–46, 52, 53, 146
2dFGRS	Two-degree-Field Galaxy Redshift Survey 8
4MOST	4-metre Multi-Object Spectroscopic Telescope 55
ACS	Advanced Camera for Surveys 106
ACT	Atacama Cosmology Telescope 8
AGN	Núcleos Activos de Galaxias (Active Galactic Nuclei) xx, xxiii, xxiv, 10–14, 16, 22, 97–111, 113–115, 117, 122–124, 127–137, 139–141, 145, 147, 149, 152, 153
ALHAMBRA	Advance Large Homogeneous Area Medium Band Redshift Astronomical xx, xxiv, 11, 13, 14, 98, 99, 102, 103, 105–111, 113–115, 117, 120, 122–124, 127–134, 136, 139, 141, 147, 152, 153
ATCA	Australia Telescope Compact Array 6
BAO	Oscilaciones Acústicas Bariónicas (Baryonic Acoustic Oscillations) xx, xxiv, 6–13, 55–58, 63, 65, 66, 68–75, 77–84, 86, 92, 93, 95–97, 146, 147, 150–152
BB	Big Bang 2, 5
BBNS	Big Bang NucleoSynthesis 2, 3, 5
BOOMERanG	Balloon Observations Of Millimetric Extragalactic Radiation and Geophysics 6
C-COSMOS	*Chandra* COSMOS-Legacy X-ray catalog 122
CDM	Cold Dark Matter 5, 6
CIC	Cloud-In-Cell 53, 57
CMB	Cosmic Microwave Background 3, 5–10
COBE	*COsmic Background Explorer* 6
COLA	COmoving Lagrangian Acceleration 57, 58, 61, 62, 64–66, 68, 69, 72, 78–81, 84–91, 95, 96

COMBO-17	Classifying Objects by Medium-Band Observations - a spectrophotometric 17-filter survey - 11, 98
COP	Centre-Of-Potential 39
COSMOS	The Cosmic Evolution Survey 11, 59, 98, 105
DEEP2	Deep Extragalactic Evolutionary Probe 2 105
DES	Dark Energy Survey 10, 20
DESI	Dark Energy Spectroscopic Instrument 10, 20, 55, 147
DM	Materia Oscura (Dark Matter) xix, xx, xxiii, xxiv, 3–5, 8–10, 12, 15–25, 27, 28, 32, 39, 41, 42, 44, 47, 48, 50, 51, 56, 57, 61, 80, 81, 84, 86, 89, 91, 145, 146, 149, 150
DMO	Dark Matter Only version of EAGLE 16, 22–29, 31–34, 42, 44, 45, 50, 51
DR	Data Release 120, 122
EAGLE	Evolution and Assembly of GaLaxies and their Environments xx, xxiv, 12, 16–18, 20–24, 26, 28–46, 48–51, 145, 146, 149, 150
ELAIS	European Large Area ISO Survey 105
ELDAR	Emission Line Detector of Astrophysical Radiators xx, xxi, xxiv, xxv, 13, 14, 98–109, 111, 113–115, 117, 122–125, 127–137, 139–141, 147, 148, 152, 153
EW	Equivalent Width 100–103, 105, 106, 109, 128–130
EdS	Einstein de Sitter 2, 3, 5, 6
FFTs	Fast Fourier Transforms 52, 53, 57
FOF	Friends-Of-Friends 22, 23
FT	Fourier Transform 52, 53, 57, 60
FWHM	Full Width Half Maximum xxiv, 56, 98, 103, 130, 132, 147
FoM	Figure of Merit 88, 89, 91–93, 147
GADGET	GAlaxies with Dark matter and Gas intEracT 57
GR	General Relativity 1, 2, 8, 9
GROTH	Deep Groth Strip Survey 105
GTC	Gran Telescopio Canarias 152
HDF-N	Hubble Deep Field North 105
HETDEX	Hobby-Eberly Telescope Dark Energy Experiment 10, 20
HOD	Halo Occupation Distribution 21, 31, 39, 40, 48, 57, 146
HSC	Subaru Hyper Suprime-Cam 10
HST	*Hubble Space Telescope* 106, 129
J-PAS	Javalambre Physics of the Accelerating Universe Astrophysical Survey 10, 11, 14, 20, 55, 56, 92, 98, 130–133, 148, 153, 154
LAS	Large Area Survey 120, 122
LSST	Large Synoptic Survey Telescope 10, 20
LePHARE	PHotometric Analysis for Redshift Estimate 99–102, 106, 109–114, 129, 130, 147
M14	Molino et al. (2014) 106, 109, 113, 114, 124
MAXIMA	Millimeter-wave Anisotropy Experiment Imaging Array 6

MCMC	Markov Chan Monte Carlo 79, 81, 84–86
MQC	Million Quasar Catalogue 110, 124
MUSE	Multi Unit Spectroscopic Explorer 150
MXXL	Millennium XXL simulation 56, 57, 83, 84, 86, 91
NFW	Navarro-Frenk-White 34, 35
PAUS	The Physics of the Accelerating Universe Survey 11, 14, 55, 98, 130, 148, 153
PDF	Probability Distribution Function 29, 41, 50, 51, 56, 59, 60, 89, 90, 99
PDZ	Redshift Probability Distribution Function 99–104, 110, 113, 129, 137, 147
PFS	Subaru Prime Focus Spectrograph 10
PSF	Point Spread Function 113, 141
RHS	Right Hand Side 59, 60, 63, 64
RSD	Redshift Space Distortions xx, xxiv, 8, 9, 11, 12, 59–62, 65, 67–69, 71–75, 78, 79, 82, 84, 92, 146
SDSS	Sloan Digital Sky Survey 8, 20, 79, 86, 105, 110, 111, 117, 120, 122, 129, 153
SED	Spectral Energy Distribution 99–103, 108, 109, 111, 113, 115, 117, 119, 124, 128, 152, 153
SFR	Star Formation Rate 149
SHAM	SubHalo Abundance Matching xx, xxiv, 12, 16–28, 30–37, 39–46, 48–50, 145, 146, 149, 150
SHARDS	Survey for High-z Absorption Red and Dead Sources 11, 98, 130, 153
SMBH	SuperMassive Black Hole 13, 97
SNR	Signal-to-Noise ratio 67–70, 75, 82, 83, 101, 130–132, 141
SPHEREx	Spectro-Photometer for the History of the universe, Epoch of Reionization, and ices Explorer 55
SPT	South Pole Telescope 8
UKIDSS	United Kingdom Infrared Telescope Infrared Deep Sky Survey 120, 122
WFIRST	Wide-Field Infrared Survey Telescope 10
WL	Weak gravitational Lensing 8–10, 145
WMAP	*Wilkinson* Microwave Anisotropy Probe 8
XMM-Newton	X-ray Multi-Mirror Mission 127
eBOSS	Extended Baryon Oscillation Spectroscopic Survey 10, 11, 20, 97, 147
zCOSMOS	zCOSMOS 10k-bright spectroscopic sample 122, 128