Mon. Not. R. Astron. Soc. **000**, 000–000 (0000) Printed 16 September 2015 (MN LATEX style file v2.2)

A Bayesian approach to linear regression in astronomy

Mauro Sereno[1⋆]
[1]*Dipartimento di Fisica e Astronomia, Università di Bologna, viale Berti Pichat 6/2, 40127 Bologna, Italia*
[2]*INAF, Osservatorio Astronomico di Bologna, via Ranzani 1, 40127 Bologna, Italia*

August 27, 2015

ABSTRACT
Linear regression is ubiquitous in astronomical analyses. I discuss a Bayesian hierarchical modeling of data with heteroscedastic and possibly correlated measurement errors and intrinsic scatter. The method fully accounts for the time evolution of the scaling. The slope, the normalization, and the intrinsic scatter of the relation can evolve with the redshift. The intrinsic distribution of the independent variable is approximated using a mixture of Gaussian distributions whose mean and standard deviation depend on time. The method can address scatter in the measured independent variable (a sort of Eddington bias), selection effects in the response variable (Malmquist bias), and departure from linearity in form of a knee. I tested the method and quantified the effect of not correcting the biases with toy models and simulations. The R-package LIRA (LInear Regression in Astronomy) is made available for performing the regression.

Key words: galaxies: clusters: general – gravitational lensing: weak

1 INTRODUCTION

Astrophysics and statistics have an interwoven history (Feigelson & Babu 2012). Linear regression is one of the most frequently used statistical techniques in astronomical data analysis. There is an impressive variety of methods to estimate functional relationships between variables.

Linear regression is kind of easy. We can draw a line which nicely interpolates a distribution of points on a paper by eye. Connecting dots and forming a regular pattern is a game fro kids. Difficulties lie in refining the results and uncovering the quantity we are really look for. As an example, the ordinary least square estimator is elegant and powerful. Still, results may be meaningless if we apply it out of its range of validity.

Most astronomical data analyses feature intrinsic scatter about the regression line. Measurement errors can affect both the independent and dependent variables. Errors may be heteroscedastic, i.e., they differ, and possibly correlated. The intrinsic distribution of the independent variables may be irregular or not uniform. The independent variable may be hidden and we could measure just a proxy of it. Selection effects can make the observed sample not representative of the population we want to study.

These aspects influence regression results and can make the use of some statistical estimators inappropriate. Many methods have been proposed to tackle these effects (Akritas & Bershady 1996; Kelly 2007; Isobe et al. 1990; Hogg, Bovy & Lang 2010; Feigelson & Babu 2012, , and references therein). Here, we are mostly interested in methods assuming that the scatters, either the intrinsic scatter or the uncertainties in the measurement process, are Gaussian. Generalized linear methods can tackle non-Gaussian multivariate datasets (de Souza et al. 2015).

Some statistical papers had the great merit to clarify the involved problematics to the astronomical community. Akritas & Bershady (1996) proposed the BCES estimator (Bivariate Correlated Errors and intrinsic Scatter) which accommodates intrinsic scatter in addition to correlated, heteroscedastic measurement errors on both variables by correcting the observed moments of the data.

Kelly (2007) described a Bayesian method (MLINMIX) based on the likelihood function of the measured data. The method can account for measurement errors, intrinsic scatter, multiple independent variables, non-detections, and selection effects in the independent variable. Kelly (2007) emphasized that the underlying distribution of covariates in a regression has to be modeled to get unbiased regression parameters and he proposed to approximate the intrinsic distribution of the independent variables as a mixture of Gaussian functions. This modeling is flexible when estimating the distribution of the true values of the independent variable and it is robust against model mispecification.

Recently, Mantz (2015) extended the MLINMIX algorithm to the case of multiple response variables and he described how to model the prior distribution of covariates using a Dirichlet process rather than a mixture. Alternative approaches based on generative models for the data were proposed too (Hogg, Bovy & Lang 2010; Robotham & Obreschkow 2015).

Here, we build upon these methods to develop a linear regression tool optimized to the study of time evolving scaling relations. Some remarkable features show up in astronomical studies.

⋆ E-mail: mauro.sereno@unibo.it (MS)

As discussed above, astronomical data sets can be affected by heteroscedatic, correlated errors in both variables and characterized by intrinsic scatter about the regression line.

Furthermore, the scaling parameters of some phenomenological relationships may not be constant with time. Time has then a special role and cannot be treated as anyone of the independent variables in a multivariate analysis. In fact, the source of intrinsic scatter is the variation in the physical properties which can be time dependent. The slope of the relation may be time dependent too if some physical processes are more conspicuous either early or lately.

In astronomical analyses, we are often interested in the correlation among an observable quantity against a variable which we do not have access to, e.g., the mass of a black hole, the mass of a galaxy cluster, the star formation rate of a galaxy. We cannot really measure these quantities but just scattered proxies of them, e.g., the weak lensing mass of a galaxy in place of the mass.

Differently from gravity, a lot of astrophysical phenomena are scale dependent. Some baryonic processes may be triggered above some thresholds and be ineffective below. This can break linearity.

Selection effects and heterogeneity can make the astronomical sample used in the regression not representative of the population we are interested in. The sample may be sparse or selected according to the value of the response variable, as in a flux limited survey.

I discuss a hierarchical Bayesian method to deal with the above aspects. Its main assumption is that scatters and uncertainties are Gaussian. Part of it has been already presented and employed in the CoMaLit (COmparing MAsses in Literature) series of papers (Sereno & Ettori 2015b, CoMaLit-I, Sereno, Ettori & Moscardini 2015, CoMaLit-II, Sereno & Ettori 2015a, CoMaLit-IV).

The method shares important features with other recently developed methods. Maughan (2014) proposed a model to constrain simultaneously the form and evolution of the scaling relations. The method distinguish between measured values, intrinsic scattered values, and model values and can constrain the intrinsic scatter and its covariance. Correlation among intrinsic scatters has to be considered in multivariate analyses to obtain unbiased scaling relations (Evrard et al. 2014; Rozo et al. 2014; Mantz et al. 2010, 2015).

If needed, I adopt the same conventions and notations of the CoMaLit series. The frame-work cosmological model is the concordance flat ΛCDM universe with density parameter $\Omega_{\rm M} = 0.3$; $H(z)$ is the redshift dependent Hubble parameter and $E_z \equiv H(z)/H_0$. 'log' is the logarithm to base 10 and 'ln' is the natural logarithm.

The method described in the present paper has been implemented in the R language[1]. The package is named LIRA (LInear Regression in Astronomy) and it is publicly available from GitHub[2] or the Comprehensive R Archive Network.

2 LINEAR SCALING

Most of the scaling relation we deal with in astronomy are time evolving power-laws. This simple schematism is supported by observations, theoretical considerations, and numerical simulations (Stanek et al. 2010; Giodini et al. 2013). The general form of the relation between two properties, e.g., the observable O and the mass M, is

$$O \propto M^{\beta} F_z^{\gamma}, \tag{1}$$

where β is the slope and the redshift evolution in the median scaling relation is accounted for by the factor F_z. According to the context, the redshift factor F_z may be either E_z or the factor $(1+z)$. In logarithmic variables, the scaling relation is linear and the scatter is Gaussian,

$$\log O = \alpha + \beta \log M + \gamma \log F_z. \tag{2}$$

In the following, $T = \log F_z$. If spectroscopically determined, measurement uncertainties in redshift are negligible[3]. The relative uncertainty in photometric redshifts can be small too, and usually smaller than relative uncertainties in other measurable properties, such as mass, luminosity, or temperature. I will not consider redshift uncertainties in the following.

In the usual framework, the time evolution does not depend on the mass scale and only affects the normalization. This is supported by the self-similar scenario, where the factor F_z for observable properties measured within the same over-density radius is E_z. However, the interplay between different physical processes that can be more or less effective at different times can make the slope time dependent, $\beta(z)$. Assuming that the evolution of the slope with redshift is linear in T, Eq. (2) can be generalized as

$$Y = \alpha + \beta\, X + \gamma\, T + \delta\, X\, T. \tag{3}$$

where $X = \log M$, and $Y = \log O$. In statisticians' terms, the critical criterion is linearity in the model parameters, not in the model variables, which makes Eq. (3) a linear model. The time variable T is deterministic, not affected by measurement errors (which I neglect). The variable X is random.

3 REGRESSION SCHEME

The Bayesian regression model presented in the following is a measurement error model (Feigelson & Babu 2012). Measurement errors are involved in a hierarchical structure and incorporated into the model. I assume that all scatter terms, i.e., intrinsic scatter and measurement errors, are Gaussian with zero mean although with different variances.

Linear regression in astronomy is usually characterized by intrinsic scatter about the scaling relation and measurement errors in both the independent and dependent variables. I assume that the covariate variable X_Z and the response variable Y_Z, which are latent, fall exactly on a straight line. This is the underlying relation we want to discover. The latent variables cannot be measured. We can measure their proxies X and Y, which differ from X_Z and Y_Z for the intrinsic scatters. These are intrinsic deviations of data points from the intrinsic scaling relation that are present even if all measurements were made with perfect precision and accuracy.

The proxies X and Y are linked to the observed manifest variables x and y with additional error terms. We could measure X and Y only in an ideal experiment with infinite accuracy and precision.

The measured values of x and y and their known measurement errors are the inputs to the model. The variables X, Y, X_Z and Y_Z have to be determined in the regression procedure. The regression scheme is summarized in Table 1 and described in details in the following.

[1] `http://www.r-project.org`
[2] `https://github.com/msereno/lira`

[3] I am not considering catastrophic errors

Table 1. Parameters of the regression scheme and their description. The variables Z is the covariate, X is a proxy of Z, and Y is the response variable. LIRA is highly customizable and priors can be easily changed. $z = z_{\rm ref}$ is the user defined reference redshift. D is either the luminosity or angular diameter distance. Default priors in square brackets have to be set by the user as delta distributions. Priors are described in Sec. 3.7.

Type	Meaning	Symbol	Code symbol	Default prior
	Y-Z scaling			
$Y_Z = \alpha_{Y\|Z} + \beta_{Y\|Z} Z + \gamma_{Y\|Z} T + \delta_{Y\|Z} Z\,T$				
Conditional scaling relation	intercept	$\alpha_{Y\|X}$	`alpha.YIZ`	`dunif`
	slope	$\beta_{Y\|X}$	`beta.YIZ`	`dt`
	time evolution	$\gamma_{Y\|Z}$	`gamma.YIZ`	`dt`
	time tilt	$\delta_{Y\|Z}$	`delta.YIZ`	`0`
$Y_Z = \alpha_{Y\|Z,\rm knee} + \beta_{Y\|Z,\rm knee} Z + \gamma_{Y\|Z,\rm knee} T + \delta_{Y\|Z,\rm knee} Z\,T$				
Scaling relation before the break	slope for $Z < Z_{\rm knee}$	$\beta_{Y\|X,\rm knee}$	`beta.YIZ.knee`	`beta.YIZ`
	time tilt for $Z < Z_{\rm knee}$	$\delta_{Y\|Z,\rm knee}$	`delta.YIZ.knee`	`delta.YIZ`
$f_{\rm knee}(Z) = 1/(1 + \exp[(Z - Z_{\rm knee})/l_{\rm knee}])$				
Transition function	break scale	$Z_{\rm knee}$	`Z.knee`	`dunif`
	break length	$l_{\rm knee}$	`l.knee`	`1e-04`
	X-Z scaling			
$X_Z = \alpha_{X\|Z} + \beta_{X\|Z} Z + \gamma_{X\|Z} T + \delta_{X\|Z} Z\,T$				
Proxy of the independent variable	bias	$\alpha_{X\|Z}$	`alpha.XIZ`	`0`
	slope	$\beta_{X\|Z}$	`beta.XIZ`	`1`
	time evolution	$\gamma_{X\|Z}$	`gamma.XIZ`	`0`
	time tilt	$\delta_{X\|Z}$	`delta.XIZ`	`0`
	Scatters			
$\sigma_{Y\|Z} = [\sigma_{Y\|Z,0} + f_{\rm knee}(Z)(\sigma_{Y\|Z,0,\rm knee} - \sigma_{Y\|Z,0})] F_z^{\gamma_{\sigma_{Y\|Z,F_z}}} D_z^{\gamma_{\sigma_{Y\|Z,D}}}$				
Intrinsic scatter	scatter at $z = z_{\rm ref}$ for $Z \geqslant Z_{\rm knee}$	$\sigma_{Y\|Z,0}$	`sigma.YIZ.0`	`prec.dgamma`
	scatter at $z = z_{\rm ref}$ for $Z < Z_{\rm knee}$	$\sigma_{Y\|Z,0,\rm knee}$	`sigma.YIZ0.knee`	`sigma.YIZ.0`
	time evolution with F_z	$\gamma_{\sigma_{Y\|Z,F_z}}$	`gamma.sigma.YIZ.Fz`	`0`
	time evolution with D	$\gamma_{\sigma_{Y\|Z,D}}$	`gamma.sigma.YIZ.D`	`0`
$\sigma_{X\|Z} = \sigma_{X\|Z,0} F_z^{\gamma_{\sigma_{X\|Z,F_z}}} D_z^{\gamma_{\sigma_{X\|Z,D}}}$				
Intrinsic scatter of the proxy	scatter at $z = z_{\rm ref}$	$\sigma_{X\|Z,0}$	`sigma.XIZ.0`	`0`
	time evolution with F_z	$\gamma_{\sigma_{X\|Z,F_z}}$	`gamma.sigma.XIZ.Fz`	`0`
	time evolution with D	$\gamma_{\sigma_{X\|Z,F_z}}$	`gamma.sigma.XIZ.D`	`0`
$\rho_{XY\|Z} = \rho_{XY\|Z,0} F_z^{\gamma_{\rho_{XY\|Z,F_z}}} D_z^{\gamma_{\rho_{XY\|Z,D}}}$				
Intrinsic scatter correlation	correlation at $z = z_{\rm ref}$	$\rho_{XY\|Z,0}$	`rho.XYIZ.0`	`0`
	time evolution with F_z	$\gamma_{\rho_{XY\|Z,F_z}}$	`gamma.rho.XYIZ.Fz`	`0`
	time evolution with D	$\gamma_{\rho_{XY\|Z,F_z}}$	`gamma.rho.XYIZ.D`	`0`
	Intrinsic distribution of the independent variable			
$p(Z) = \sum_k \pi_k \mathcal{N}(\mu_{Z,k}(z), \sigma_{Z,k}(z))$				
Gaussian mixture	number of components	$n_{\rm mix}$	`n.mixture`	`[1]`
	weights of the components	π_k	`pi[k]`	`ddirch`
$\mu_{Z,k}(z) = \mu_{Z,0k} + \gamma_{\mu_Z,F_z} T + \gamma_{\mu_Z,D} \log D$				
Means of the Gaussian distributions	mean of the first component at $z = z_{\rm ref}$	$\mu_{Z,01}$	`mu.Z.0`	`dunif`
	means of the additional components	$\mu_{Z,0k}$	`mu.Z.0.mixture[k]`	`dunif`
	time evolution with F_z	γ_{μ_Z,F_z}	`gamma.mu.Z.Fz`	`dt`
	time evolution with D	$\gamma_{\mu_Z,D}$	`gamma.mu.Z.D`	`dt`
$\sigma_{Z,k}(z) = \sigma_{Z,0k} F_z^{\gamma_{\sigma_Z,F_z}} D_z^{\gamma_{\sigma_Z,D}}$				
Standard deviations of the Gaussian distributions	deviation of the first component at $z = z_{\rm ref}$	$\sigma_{Z,01}$	`sigma.Z.0`	`prec.dgamma`
	deviation of the additional components ($2 \leqslant k \leqslant n_{\rm mix}$) at $z = z_{\rm ref}$	$\sigma_{Z,0k}$	`sigma.Z.0.mixture[k]`	`prec.dgamma`
	time evolution with F_z	γ_{σ_Z,F_z}	`gamma.sigma.Z.Fz`	`0`
	time evolution with D	$\gamma_{\sigma_Z,D}$	`gamma.sigma.Z.D`	`0`

3.1 Linear scaling

The linear relation between two unscattered quantities (I am not counting the time) can be expressed as

$$Y_Z = \alpha_{Y|Z} + \beta_{Y|Z} Z + \gamma_{Y|Z} T + \delta_{Y|Z} Z\, T, \tag{4}$$

where α denotes the normalization, the slope β accounts for the dependence with Z, the slope γ accounts for the time-evolution of the normalization and δ quantifies the tilt of the slope with time.

The basic case summarized in Eq. (4) is enough to describe the regression of Y_Z against a variable which is directly observable. This is the case of the luminosity versus temperature relation of galaxy clusters. In some other cases, the independent variable Z is not directly available from measurement. For example, we cannot measure the mass of a cluster (Z), but we can approximate it with the weak lensing mass (X). We have then to couple Eq. (4) with

$$X_Z = \alpha_{X|Z} + \beta_{X|Z} Z + \gamma_{X|Z} T + \delta_{X|Z} Z\, T, \tag{5}$$

In this case, X_Z and Y_Z are related to the same covariate variable, Z. The relations among Z, X_Z and Y_Z are deterministic and they are not affected by scatter. X_Z and Y_Z are rescaled versions of the latent variable Z, which can be seen as a fundamental property of the object, e.g., the mass of a cluster of galaxies.

3.2 Measurement uncertainties

The measured quantities $\boldsymbol{x}$ and $\boldsymbol{y}$ are the manifest values of $\boldsymbol{X}$ and $\boldsymbol{Y}$[4]. Due to observational uncertainties their relation can be expressed as

$$P(\boldsymbol{x}_i, \boldsymbol{y}_i | \boldsymbol{X}_i, \boldsymbol{Y}_i) = \mathcal{N}^{\rm 2D}(\{\boldsymbol{X}_i, \boldsymbol{Y}_i\}, \mathbf{V}_{\delta,i}), \tag{6}$$

where $\mathcal{N}^{\rm 2D}$ is the bivariate Gaussian distribution and $\mathbf{V}_{\delta,i}$ is the uncertainty covariance matrix whose diagonal elements are denoted as $\delta^2_{x,i}$ and $\delta^2_{y,i}$, and whose off-diagonal elements are denoted as $\rho_{xy,i}\delta_{x,i}\delta_{y,i}$.

As a result of the i-th measurement process, we obtain $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}$ and the related uncertainty covariance matrix $\mathbf{V}_{\delta,i}$. The variables $\boldsymbol{X}_{Z,i}$, $\boldsymbol{Y}_{Z,i}$, $\boldsymbol{X}_i$, and $\boldsymbol{Y}_i$, are unknown variables to be determined under the assumption of linearity.

3.3 Intrinsic scatter

The intrinsic scatter quantifies how close the data distribution is to linearity. The true properties of an astronomical object X and Y, which we can try to measure, are intrinsically scattered with respect to the latent model variables X_Z and Y_Z, which fall on a line without deviations but which are hidden properties.

Observable properties are usually log-normally distributed about the mean scaling relations (Stanek et al. 2010; Fabjan et al. 2011; Angulo et al. 2012; Saro et al. 2013). This is supported by numerical simulations (Stanek et al. 2010; Fabjan et al. 2011; Angulo et al. 2012) and observational studies (Maughan 2007; Vikhlinin et al. 2009). We assume that the intrinsic scatters are Gaussian,

$$P(\boldsymbol{X}_i, \boldsymbol{Y}_i | \boldsymbol{X}_{Z_i}, \boldsymbol{Y}_{Z,i}) = \mathcal{N}^{\rm 2D}(\{\boldsymbol{X}_{Z,i}, \boldsymbol{Y}_{Z,i}\}, \mathbf{V}_{\sigma,i}), \tag{7}$$

where $\mathbf{V}_{\sigma,i}$ is the scatter covariance matrix of the i-th cluster whose diagonal elements are denoted as $\sigma^2_{X|Z,i}$ and $\sigma^2_{Y|Z,i}$, and whose off-diagonal elements are denoted as $\rho_{XY|Z,i}\sigma_{X|Z,i}\sigma_{Y|Z,i}$.

[4] $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{X}$, and $\boldsymbol{Y}$ are vectors of n elements.

The intrinsic scatter of a scaling relation is related to the degree of regularity of the sample. The scatter can be prominent in morphologically complex halos or in objects which depart from dynamical/hydrostatic equilibrium (Fabjan et al. 2011; Saro et al. 2013). Deviations from spherically symmetry are another major source of scatter (Limousin et al. 2013; Sereno et al. 2013). Since high redshift objects are more irregular and less spherical, the scatter is usually expected to increase with redshift Saro et al. (2013); Fabjan et al. (2011). The degree of scatter and its evolution depends on the baryonic physics too (Fabjan et al. 2011).

The time evolution of the scatters and of their correlation can be modeled as (CoMaLit-IV)

$$\sigma_{X|Z}(z) = \sigma_{X|Z,0} F_z^{\gamma_{\sigma_{X|Z,F_z}}} D_z^{\gamma_{\sigma_{X|Z,D}}}, \tag{8}$$

$$\sigma_{Y|Z}(z) = \sigma_{Y|Z,0} F_z^{\gamma_{\sigma_{Y|Z,F_z}}} D_z^{\gamma_{\sigma_{Y|Z,D}}}, \tag{9}$$

$$\rho_{XY|Z}(z) = \rho_{XY|Z,0} F_z^{\gamma_{\rho_{XY|Z,F_z}}} D_z^{\gamma_{\rho_{XY|Z,D}}}, \tag{10}$$

where D_z is either the luminosity or the angular diameter distance.

If we want to regress Y against X, we can identify X and Z. Equation (7) reduces to

$$P(\boldsymbol{Y}_i | \boldsymbol{Z}_i) = \mathcal{N}(\boldsymbol{Y}_{Z,i}, \sigma^2_{Y|Z,i}), \tag{11}$$

3.4 Malmquist bias

Selection effects are a common concern in the astronomical analysis. If only objects above an observational threshold (in the response variable) are included, the sample is affected by the Malmquist bias. In this case, the relation between the measured and the true values (Eq. 6) or between the true values and the unscattered values (Eq. 7) are modified.

The bias can be modeled by truncating the probability distributions below the threshold $\boldsymbol{y}_{\rm th,i}$. The measured and the true values of the quantities are now related as

$$P(\boldsymbol{x}_i, \boldsymbol{y}_i | \boldsymbol{X}_i, \boldsymbol{Y}_i) \propto \mathcal{N}^{\rm 2D}(\{\boldsymbol{X}_i, \boldsymbol{Y}_i\}, \mathbf{V}_{\delta,i}) \mathcal{U}(\boldsymbol{y}_{\rm th,i},), \tag{12}$$

where $\mathcal{U}$ is the uniform distribution null for $y < \boldsymbol{y}_{\rm th,i}$.

The observational thresholds $\boldsymbol{y}_{\rm th}$ may not be exactly known. This may be the case when the quantity which the selection procedure is based on differs from the quantity used in the regression. We have then to consider the additional relation

$$P(\boldsymbol{y}_{\rm th,i}) = \mathcal{N}(\boldsymbol{y}_{\rm th,obs,i}, \delta^2_{y_{\rm th,i}}), \tag{13}$$

where $\delta_{y_{\rm th,i}}$ is the uncertainty associated to the measured threshold $\boldsymbol{y}_{\rm th,obs,i}$. Equations (12 and 13) can be combined by considering a sigmoid curve instead of the step function in Eq. (12).

The conditional probability of the proxies in the sample is truncated too,

$$P(\boldsymbol{X}_i, \boldsymbol{Y}_i | \boldsymbol{X}_{Z,i}, \boldsymbol{Y}_{Z,i}) = \mathcal{N}^{\rm 2D}(\{\boldsymbol{X}_{Z,i}, \boldsymbol{Y}_{Z,i}\}, \mathbf{V}_{\sigma,i}) \mathcal{U}(\boldsymbol{Y}_{\rm th,i},), \tag{14}$$

where the threshold $\boldsymbol{Y}_{\rm th,i}$ follows the distribution

$$P(\boldsymbol{Y}_{\rm th,i}) = \mathcal{N}(\boldsymbol{y}_{\rm th,i}, \delta^2_{y_i}). \tag{15}$$

3.5 Intrinsic distribution

The proper modeling of the distribution of the independent variable is crucial. Samples considered in regression analyses are usually biased with respect to the parent population. Sources may be selected according to their properties. Furthermore, even in absence of selection effect the intrinsic parent population is usually not uniform, which may cause tail effects.

The intrinsic distribution of the independent variable Z is shaped by two main effects. On one hand, very massive objects are rarer. On the other hand, massive objects are usually strong emitters and are easier to be detected to very large distances. As a result, the shape of the distribution is fairly unimodal and it evolves with time (CoMaLit-IV).

The combined evolution of the completeness and of the parent population can be characterized through the evolution of the peak and of the dispersion of the distribution of the selected sample. The intrinsic distribution of Z can be approximated with a mixture of $n_{\rm mix}$ time-evolving Gaussian functions (Kelly 2007, CoMaLit-II, CoMaLit-IV),

$$p(Z) = \sum_{k=1}^{n_{\rm mix}} \pi_k \, \mathcal{N}\left(\mu_{Z,k}(z), \sigma^2_{Z,k}(z)\right), \tag{16}$$

where π_k is the probability of drawing a data point from the k-th component, $\sum_k \pi_k = 1$.

I assume that the mixture components have different mean and dispersion but share the same evolution parameters. The mean of each component is connected to the (redshift-evolving) observational thresholds and to the intrinsic scatter of the observable quantity used to select the clusters, which evolves too. As a result, the evolution of the (mean of the) k-th mixture component can be modeled as (CoMaLit-IV),

$$\mu_{Z,k}(z) = \mu_{Z,0k} + \gamma_{\mu_Z,F_z} T + \gamma_{\mu_Z,D} \log D_z, \tag{17}$$

where $\mu_{Z,0k}$ is the mean at the reference redshift.

The evolution of the dispersions is related to the intrinsic scatter of the observable property used to select the samples. The time dependence can be modeled as (CoMaLit-IV)

$$\sigma_{Z,k}(z) = \sigma_{Z,0k} F_z^{\gamma_{\sigma_Z,F_z}} D_z^{\gamma_{\sigma_Z,D}}. \tag{18}$$

The proper modeling of the intrinsic distribution of the independent variable is crucial to correct for the Eddington bias, when the average value of an observed sample differs from the true intrinsic average of the objects of the same class (Eddington 1913; Jeffreys 1938; Eddington 1940;CoMaLit-I).

3.6 Departure from linearity

Physical processes are effective at different scales, which may cause deviation from linearity. Gravity is the driving force behind formation and evolution of galaxy clusters but at small scales baryonic physics can play a prominent role. As a result, linearity can break. This can be shaped with a knee in the relation, such that before the breaking scale $Z_{\rm knee}$, the scaling follows

$$Y_Z = \alpha_{Y|Z,\rm knee} + \beta_{Y|Z,\rm knee} Z + \gamma_{Y|Z,\rm knee} T + \delta_{Y|Z,\rm knee} Z\, T. \tag{19}$$

The normalization $\alpha_{Y|Z,\rm knee}$ and the time evolution $\gamma_{Y|Z,\rm knee}$ are determined by requiring equality at the transition $Z_{\rm knee}$,

$$\alpha_{Y|Z,\rm knee} = \alpha_{Y|Z} + (\beta_{Y|Z} - \beta_{Y|Z,\rm knee}) \tag{20}$$

$$\gamma_{Y|Z,\rm knee} = \gamma_{Y|Z}. \tag{21}$$

The transition between the two regimes can be modeled through a S-shaped transition function,

$$f_{\rm knee} = \frac{1}{1+\exp\left[(Z - Z_{\rm knee})/l_{\rm knee}\right]}, \tag{22}$$

where the scale $l_{\rm knee}$ sets the transition length. The relation over the full range reads

$$Y_Z = \alpha_{Y|Z} + \beta_{Y|Z} Z + \gamma_{Y|Z} T + \delta_{Y|Z} Z\, T + (Z_{\rm knee} - Z) f_{\rm knee}(Z) \times \left[(\beta_{Y|Z} - \beta_{Y|Z,\rm knee}) + (\delta_{Y|Z} - \delta_{Y|Z,\rm knee})\, T\right], \tag{23}$$

The same physical processes can affect the scatter too, which I model as

$$\sigma_{Y|Z}(Z, z_{\rm ref}) = \sigma_{Y|Z,0} + (\sigma_{Y|Z,0,\rm knee} - \sigma_{Y|Z,0}) f_{\rm knee}(Z). \tag{24}$$

I assume that the redshift evolution of the scatter is not affected.

3.7 Priors

The Bayesian statistical treatment requires the explicit declaration of the priors. Priors can be either conveniently non-informative, if we have no guess on the parameters (CoMaLit-I; CoMaLit-II), or peaked and with small dispersion, to convey the information obtained with exiting experiments or theory. The LIRA approach allows the user to set the prior distribution of all parameters. The parameters can be also frozen by fixing them with a delta-prior.

Standard priors on the intercept $\alpha_{Y|Z}$ and on the means of the mixture components $\mu_{Z,0,k}$ can be flat,

$$\alpha_{Y|Z},\ \mu_{Z,0,k} \sim \mathcal{U}(-n_{\rm L}, n_{\rm L}), \tag{25}$$

where n_L is a large number. In LIRA, the default value is $n_{\rm L} = 10^4$ and the shortcut for the prior is `dunif`.

The slopes can follow the Student's t_1 distribution with one degree of freedom, as suitable for uniformly distributed direction angles,

$$\beta_{Y|Z},\ \gamma_{Y|Z},\ \gamma_{\mu_Z,E_z},\ \gamma_{\mu_Z,D} \sim t_1. \tag{26}$$

In LIRA, the shortcut for this prior is `dt`. The γ-type parameters are set to zero when no redshift information is provided. The other slope parameters ($\beta_{X|Z}$, the δ's, the other γ's) are by default frozen to 0. They can be unpegged by setting other priors. In these cases, the non-informative t_1 prior is suggested.

For the variance, I adopted by default a scale-invariant scaled inverse χ^2-distribution,

$$\sigma^2_{Y|Z,0},\ \sigma^2_{Z,0,k} \sim \text{Scale-inv-}\chi^2(\nu, \xi), \tag{27}$$

with $\nu = 2/n_{\rm L}$ degrees of freedom and scale $\xi = 1$[5].

By default, X tallies Z and it is unscattered, $\sigma_{X|Z,0} = 0$. The scatter correlation $\rho_{XY|Z,0}$ is set to zero too. Otherwise, a flat prior can be adopted,

$$\rho_{XY|Z,0} \sim \mathcal{U}(-1, 1). \tag{28}$$

The parameters in the scaling $Y - Z$ and $X - Z$, see Eqs. (4 and 5) are redundant. If we do not know the value of Z, we cannot measure all of them. By default, I assumed that X is an unbiased proxy of Z, i.e., $\alpha_{X|Z} = 0$, $\beta_{X|Z} = 1$, $\gamma_{X|Z} = 0$, and $\delta_{X|Z} = 0$. For linear relations, fixing the parameters of the X-Z rather than the Y-Z relation is just a matter of rescaling. In absence of a direct measurement of Z, the bias between X and Z (i.e, $\alpha_{X|Z} \neq 0$) is degenerate with the estimated overall normalization of the scaling between Y and Z. The regression can only constrain the relative bias between X and Y (CoMaLit-I).

[5] The inverse χ^2 prior for the variance is equivalent to a Gamma distribution $\Gamma(r = 1/n_{\rm L}, \lambda = 1/n_{\rm L})$ for the precision, i.e., the inverse of the variance. Hence, the name `prec.dgamma` for this prior in LIRA. `prec.dgamma` is the only LIRA prior which models the variance. Other customizable priors refer directly to the standard deviations.

By default, I considered a single Gaussian distribution to model the intrinsic distribution of the independent variable ($n_{\rm mix} = 1$). For mixtures, I adopted a Dirichlet distribution for the probability coefficients (Kelly 2007)

$$\pi_1, ..., \pi_{n_{\rm mix}} \sim \text{Dirichlet}(1, ..., 1), \tag{29}$$

which is equivalent to a uniform prior under the constraint $\sum_{k=1}^{n_{\rm mix}} \pi_k = 1$. In LIRA, the shortcut for this prior is `ddirch`. The number of mixture components $n_{\rm mix}$ has to be fixed. Alternative approaches can determine the optimal number of Gaussian components modeling the intrinsic distribution through a Dirichlet process (Mantz 2015).

By default, the regression adopts linear models with no breaks. There is no knee and the slope $\beta_{Y|Z,\rm knee}$ and tilt $\delta_{Y|Z,\rm knee}$ tally $\beta_{Y|Z}$ and $\delta_{Y|Z}$, respectively. Otherwise, a flat prior can be adopted for $Z_{\rm knee}$ and a Student's-t prior for $\beta_{Y|Z,\rm knee}$ and $\delta_{Y|Z,\rm knee}$, when applicable. The transition length is set by default to $l_{\rm knee} = 10^{-4}$.

The above motivations drove the choice of the default priors listed in Table 1. For current data-sets, the γ_D parameters can be set safely to zero in any regression, see App. B. The only exception is $\gamma_{\mu_Z,D}$, which is crucial to model the time-evolution of the intrinsic function of flux-selected samples (CoMaLit-IV).

4 SIMULATIONS

I investigated how the approach detailed in Sec. 3 can recover scaling relations in presence of noise, scatter, and selection biases. The approach was tested with toy models and simulated data samples. I set up a basic scheme, which was modified if needed to highlight some aspects. The essential features were as follows.

The independent variables Z were drawn from a normal distribution with mean $\mu_{Z,0} = 0$, and standard deviation $\sigma_{Z,0} = 0.3$. The values of Y were simulated assuming $\alpha_{Y|Z} = 0$ and $\beta_{Y|Z} = 1$ and $\sigma_{Y|Z,0} = 0.1$. X tallies with Z ($\alpha_{X|Z} = 0$ and $\beta_{X|Z} = 1$ and $\sigma_{X|Z,0} = 0.0$). All other parameters were set to zero by default.

The measurement errors were different for each data point. The variances in the measurement errors, δx^2 and δy^2 were drawn from a scaled inverse χ^2-distribution with 5 degrees of freedom (Kelly 2007). The scale parameters, which dictates the typical size of the measurements errors were set to 0.1^2. I simulated a varying degree of correlations among the measurement errors. The correlations were drawn from a uniform distribution ranging from 0.0 to 0.4.

In case of samples covering a redshift range, the above parameters values were intended as the normalizations at the reference redshift, $z_{\rm ref} = 0.01$. The default time evolution was set to $\gamma_{Y|Z} = 1$ and the redshifts were drawn from a lognormal distribution. I considered $F_z = E_z$ as time factor and I computed cosmological distances as angular diameter distances.

For each case study I generated 10^3 data sets, each one with $n_{\rm sample} = 100$ data points, as typical of current samples (CoMaLit-IV). The scaling relations, the scatters, and the intrinsic Z-distributions were recovered with LIRA. Parameter priors were set to the default distributions listed in Table 1. Posterior probability distributions were constrained with Markov chains generated with a Gibbs sampler. The LIRA package relies on JAGS (Just Another Gibbs sampler) library to perform the sampling[6].

[6] JAGS is publicly available at `http://mcmc-jags.sourceforge.net`.

Table 2. Scaling parameters recovered from samples whose independent variable follows a skewed distribution. I report the bi-weight estimators of the distribution of the median values of the simulated chains.

parameter	input	LIRA $n_{\rm mix} = 1$	LIRA $n_{\rm mix} = 3$	MLINMIX $n_{\rm mix} = 3$	BCES
			$z = z_{\rm ref}$		
$\alpha_{Y\|Z}$	[0]	0.00 ± 0.03	0.00 ± 0.03	0.00 ± 0.03	0.00 ± 0.04
$\beta_{Y\|Z}$	[1]	1.02 ± 0.12	1.01 ± 0.11	1.00 ± 0.11	1.02 ± 0.15
$\sigma_{Y\|Z,0}$	[0.1]	0.09 ± 0.03	0.09 ± 0.03	0.10 ± 0.02	0.13 ± 0.01
			redshift evolution		
				$n_{\rm mix} = 1$	
$\alpha_{Y\|Z}$	[0]	-0.01 ± 0.10		0.01 ± 0.09	
$\beta_{Y\|Z}$	[1]	1.02 ± 0.12		0.99 ± 0.11	
$\gamma_{Y\|Z}$	[1]	0.83 ± 0.45		1.02 ± 0.47	
$\sigma_{Y\|Z,0}$	[0.1]	0.09 ± 0.03		0.10 ± 0.02	

For each data set, I computed the parameter medians from the chains and I studied the distributions of the medians of the ensemble.

The simulation scheme was modified if needed to highlight some aspects. On occasion, I simulated a skewed and evolving intrinsic distribution of the independent variables, scattered values of X, time evolving scatter and slope.

When applicable, I also considered other publicly available methods such as BCES[7] (Akritas & Bershady 1996) and LINMIX[8] or its generalization to multivariate regression MLINMIX[9] (Kelly 2007). The underlying hypotheses of these methods are well known and I only used them when applicable. I did not consider BCES for time-dependent populations and MLINMIX for time evolving scatters, Malmquist biased samples or in case of deviation from linearity.

4.1 Skewed distribution

The accurate modeling of the intrinsic distribution of the covariate variable is crucial to unbiased linear regression. I considered an asymmetric distribution. I modified the basic simulation scheme by drawing Z from a skew-normal distribution with shape parameter $\alpha_{Z,0,\rm skew} = 3.0$. The location and the scale parameter were made to coincide with the mean and the standard deviation of the basic normal distribution. Results are reported in Table 2.

4.1.1 No time evolution

I first considered samples drawn at the same reference redshift. Results are summarized in Table 2 and Fig. 1. To recover the parameters, I considered either a simple LIRA model with just one normal distribution to shape $p(Z)$ or a mixture of three components. For comparison, I also computed parameter chains with LINMIX adopting a mixture of three Gaussian distributions and the BCES(Y|X) estimator. The original work introducing BCES did not advocate any method to compute the intrinsic scatter, which I computed following Pratt et al. (2009).

Input parameters are well reproduced by all methods. In this

[7] `http://www.astro.wisc.edu/~mab/archive/stats/stats.html`
[8] `http://idlastro.gsfc.nasa.gov/ftp/pro/math/linmix_err.pro`
[9] `http://idlastro.gsfc.nasa.gov/ftp/pro/math/mlinmix_err.pro`

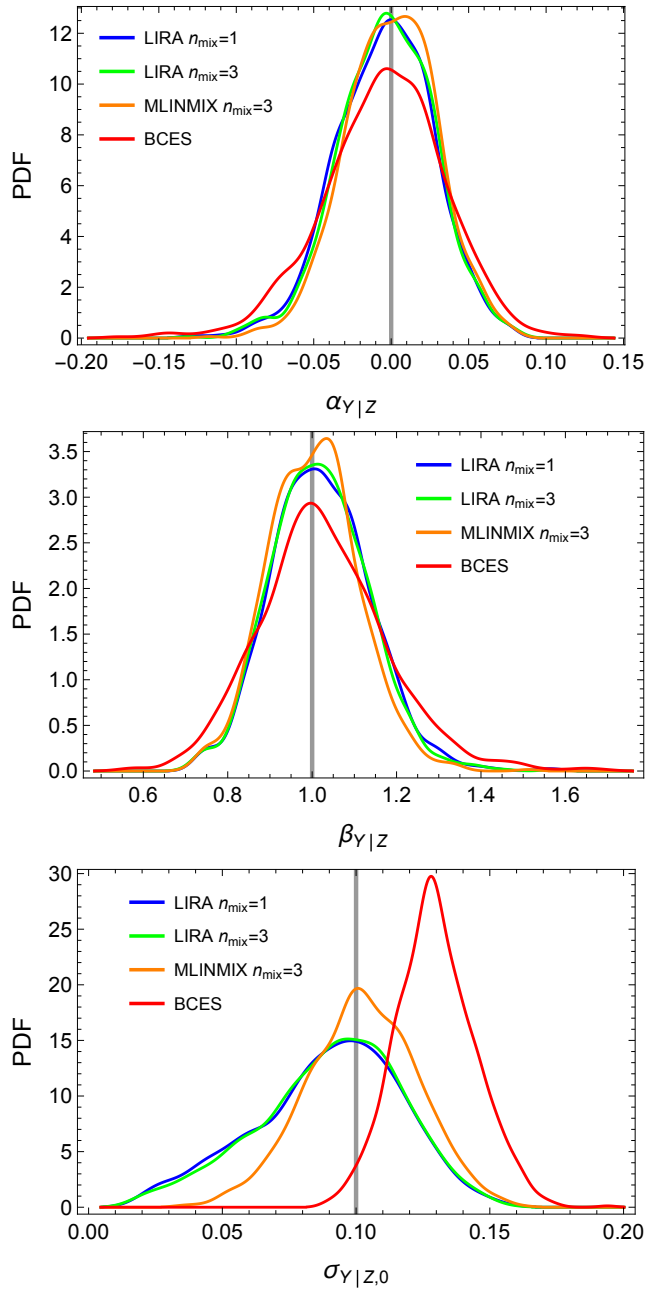

Figure 1. Distribution of the median parameters obtained from samples drawn from a skewed intrinsic distribution $p(Z)$. The blue, green, orange and red lines are the smoothed histogram of the distributions obtained with LIRA by modeling $p(Z)$ with a single Gaussian function, LIRA by adopting a mixture of 3 Gaussian functions, LINMIX by adopting a mixture of 3 Gaussian functions, and BCES. The vertical gray lines are set at the values of the input parameters. From the top to the bottom panel: the intercept, the slope and the intrinsic scatter.

setting, the agreement between LIRA and LINMIX is excellent. This is expected since the main assumptions of the two methods are equivalent. Minor differences comes from the slightly different priors. The parameter distributions agree very well, even though the distribution of the intrinsic scatter $\sigma_{Y|Z,0}$ from LIRA has a more pronounced tail at small values.

BCES recovers well the central values of the slope and of the intercept but statistical uncertainties are larger. The intrinsic scatter estimate is biased high.

Even though the intrinsic simulated distribution of Z is skewed, there is no real improvement by augmenting the number of mixture components, see Table 2. As far as the intrinsic distribution is unimodal and the sample is not too rich, one Gaussian component is enough.

4.1.2 Evolution with redshift

I considered samples with covering an extended redshift range. Redshifts were drawn from a lognormal distribution such that $\ln z$ has mean $\ln(0.3)$ and standard deviation 0.5. In these simulations, the skewed distribution of the independent variable is time evolving. The location parameter of the input distribution evolves with the redshift as in Eq. (17) with $\gamma_{\mu_Z,F_z} = 0.5$ and $\gamma_{\mu_Z,D} = 0.5$; the scale parameter is fixed, whereas the shape parameter evolves with redshift proportionally to E_z. The input intrinsic scatter $\sigma_{Y|Z}$ is redshift independent.

I recovered the input parameters by modeling $p(Z)$ with a single normal distribution whose mean and standard deviation evolve with time. The prior on γ_{σ_Z,F_z} was set such that the inverse variance follows a Gamma distribution, see Eq. (27), whereas $\gamma_{\sigma_Z,D}$ was set to zero. Even if the input $p(Z)$ distribution is asymmetric, this modeling is enough to recover the time evolution of $p(Z)$, see Figure 2, and to get unbiased values of the scaling parameters, see Table 2 and Figure 3.

For comparison I performed the regression with MLINMIX too. The LIRA scheme differs from the multivariate analysis detailed in Kelly (2007) in one major feature. LIRA models the intrinsic distribution with a mixture of one-dimensional Gaussian components whose means and standard deviations are time-dependent. On the other hand, MLINMIX models the bi-dimensional distribution of Z and T with a mixture of bi-dimensional Gaussian components whose means and variances are not time-evolving. Notwithstanding this important difference and some minor differences due to the prior choice, both approaches can recover with good accuracy the scaling parameters, see Figure 3.

As far as the scaling parameters and the scatter is concerned, it is important to model the non-uniformity of the distribution of the intrinsic distribution. Details on the exact form of the distribution are of second order. The safer approach to model noisy ad sparse samples is to use the simplest model, e.g., a single normal distribution for $p(Z)$. In samples of order of one hundred of objects, there are just a few items at high redshift. Enforcing a more complex distribution, such as a skewed Gaussian, to model sparse data can bias the results toward a few outliers due to overfitting. Complex distributions are recommended only for very rich samples.

4.2 Eddington bias

The Eddington bias can affect the estimate of the scaling parameters whether the measurement errors on the covariate variable are not accounted for (Eddington 1913; Jeffreys 1938; Eddington 1940) or the X variable is a scattered proxy of the latent covariate Z (Sereno & Ettori 2015b). Here, we are mostly interested in the second case, which is often overlooked.

I simulated the samples by assuming that the measurable X is an unbiased ($\alpha_{X|Z} = 0$ and $\beta_{X|Z} = 1$) but scattered ($\sigma_{X|Z,0} = 0.1$) proxy of Z. Results are summarized in Table 3 and Fig. 4.

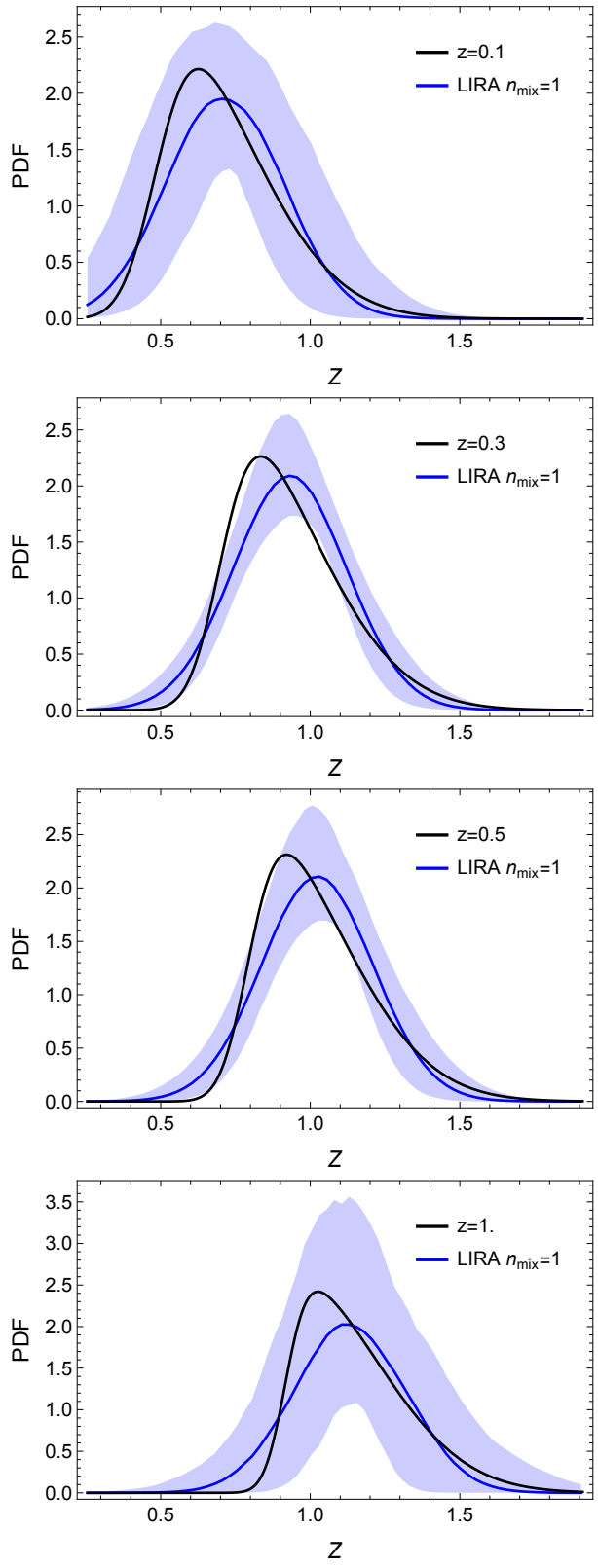

Figure 2. The reconstructed intrinsic distribution of the independent variable Z at different redshifts. From top to the bottom, $z = 0.1, 0.3, 0.5, 1.0$. The black line is the input distribution, the blue line is the median reconstructed relation, the shadowed blue region encloses the 1-σ confidence region for each value of Z. For a total of $n_{\rm sample} = 100$ data, we expect ~ 21, 51, 20 and 1 sources in the redshift range $0.0 \leqslant z \leqslant 0.2$, $0.2 \leqslant z \leqslant 0.4$, $0.4 \leqslant z \leqslant 0.6$ and $0.9 \leqslant z \leqslant 1.1$, respectively.

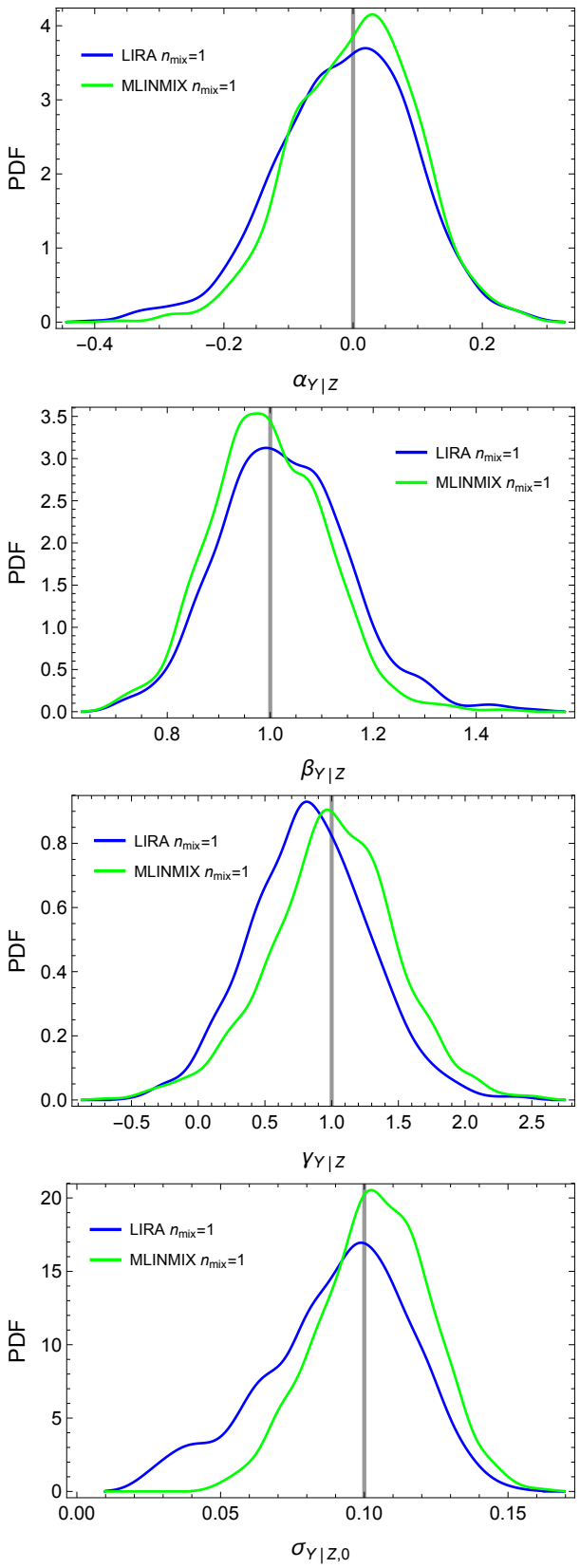

Figure 3. Distribution of the median parameters obtained from samples with a skewed and time-evolving intrinsic distribution $p(Z, z)$. The blue (green) line is the smoothed histogram of the distribution obtained with LIRA (MLINMIX). $p(Z, z)$ was modeled with a single Gaussian function. The vertical gray lines are set at the values of the input parameters. From the top to the bottom panel: the intercept, the slope, the time evolution, and the intrinsic scatter.

Table 3. Scaling parameters recovered from biased samples. For each case (listed in Col. 1), I reported on consecutive rows the values of the parameters obtained with regressions which do either correct or not for the bias. Reported values are the bi-weight estimators of the distribution of the median values of the simulated chains.

case	$\alpha_{Y\|Z}$	$\beta_{Y\|Z}$	$\sigma_{Y\|Z,0}$
input	[0]	[1]	[0.1]
Eddington bias ($\sigma_{X\|Z} \neq 0$)			
corrected	0.00 ± 0.02	1.01 ± 0.09	0.09 ± 0.03
biased	0.00 ± 0.02	0.90 ± 0.07	0.14 ± 0.02
Malmquist bias			
corrected	0.01 ± 0.03	1.06 ± 0.11	0.09 ± 0.03
biased	0.03 ± 0.02	0.90 ± 0.09	0.08 ± 0.03
Linearity break (knee)			
corrected	0.00 ± 0.03	0.98 ± 0.18	0.10 ± 0.04
biased	-0.05 ± 0.02	1.31 ± 0.12	0.15 ± 0.03

The induced bias is significant. The Eddington bias makes the observed relation flatter and inflates the intrinsic scatter. Since I considered a scatter $\sigma_{X|Z}$ independent of Z, the bias has a symmetric action and the pivot point of the relation does not change. The normalization is not affected. Statistical uncertainties on the regression parameters are underestimated, as usual in biased measurements.

4.3 Malmquist bias

The Malmquist bias has long been known (Malmquist 1920). Still, it can be difficult to tackle. Proposed recipes consider the correction of the measured values of individual objects, which needs a guess on the intrinsic scatter, or the modeling through a proper definition of the selection efficiency in the likelihood function (Vikhlinin et al. 2009).

I simulated the samples as in the standard case but I only kept objects whose measured response variable y exceeded a threshold value ($y > y_{\rm th} = -0.3$). Nearly 80 per cent of the items makes the cut. Results are summarized in Table 3 and Fig. 5.

The Malmquist bias make the observed relation flatter. If the bias is not corrected for, the measured slope is biased toward 0 whereas the measured intercept is biased high. The scatter is affected too. It can be underestimated,

4.4 Linearity break

The scales and the effectiveness of different physical processes can break the linearity of a scaling relation. In the formation and evolution of galaxy clusters, baryonic and energetic effects are relevant in small objects and can challenge the dominance of the gravitational force. A bent scaling relation can be more apt to model the process.

I simulated a broken power law relation. I set the knee at $Z_{\rm knee} = \mu_{Z,0} - \sigma_{Z,0}$, i.e., $\sim$ 16 per cent of the sources follow a different scaling. The slope before the break was set at $\beta_{Y|Z,{\rm knee}} = 3.0$. Results are summarized in Table 3 and Fig. 6.

Parameter estimates obtained with a simple linear model are severely biased. Without a knee, the model cannot distinguish the two regimes and the measured slope is a weighted mean of the two real slopes. Being the slope before the knee steeper in the simulation, the intercept estimated by the linear model is biased low. If not modeled, the knee strongly affects the estimated scatter. To mimic the break and the steeper slope, the estimated scatter is severely overestimated.

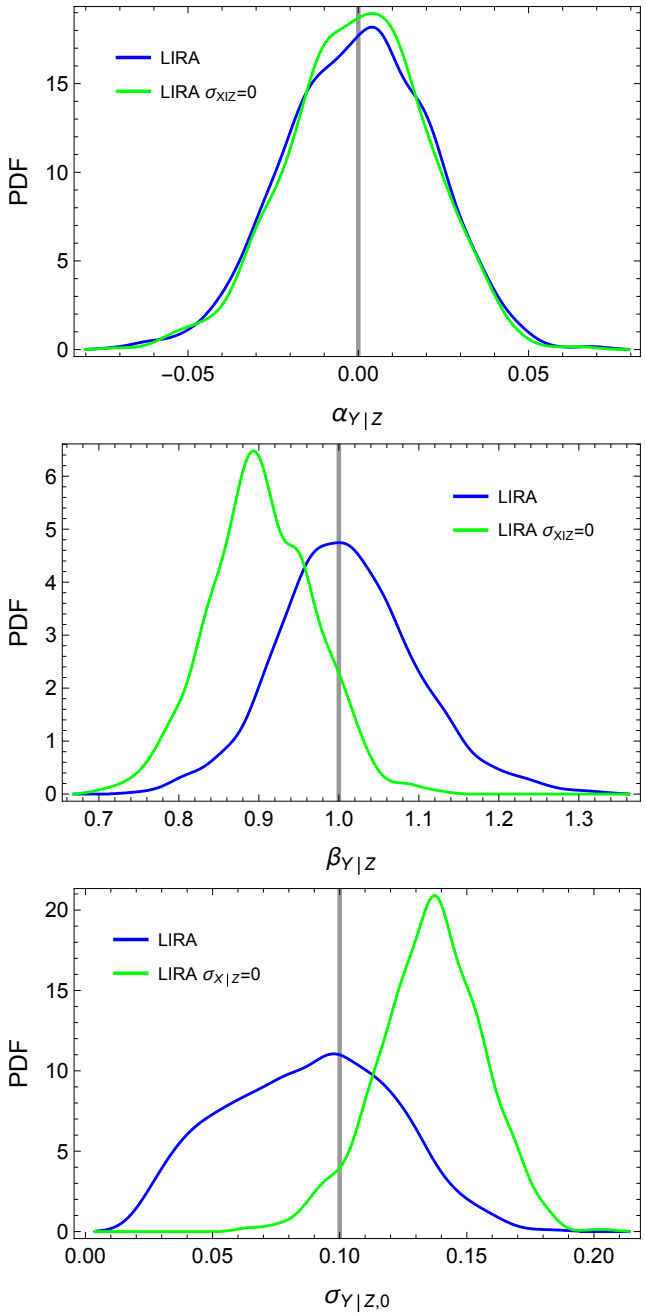

Figure 4. Distribution of the median parameters obtained from samples affected by Eddington bias. The blue line is the smoothed histogram of the distribution obtained by considering the intrinsic scatter in the covariate variable. The green line plots the results from a biased fit. The vertical gray lines are set at the values of the input parameters. From the top to the bottom panel: the intercept, the slope and the intrinsic scatter.

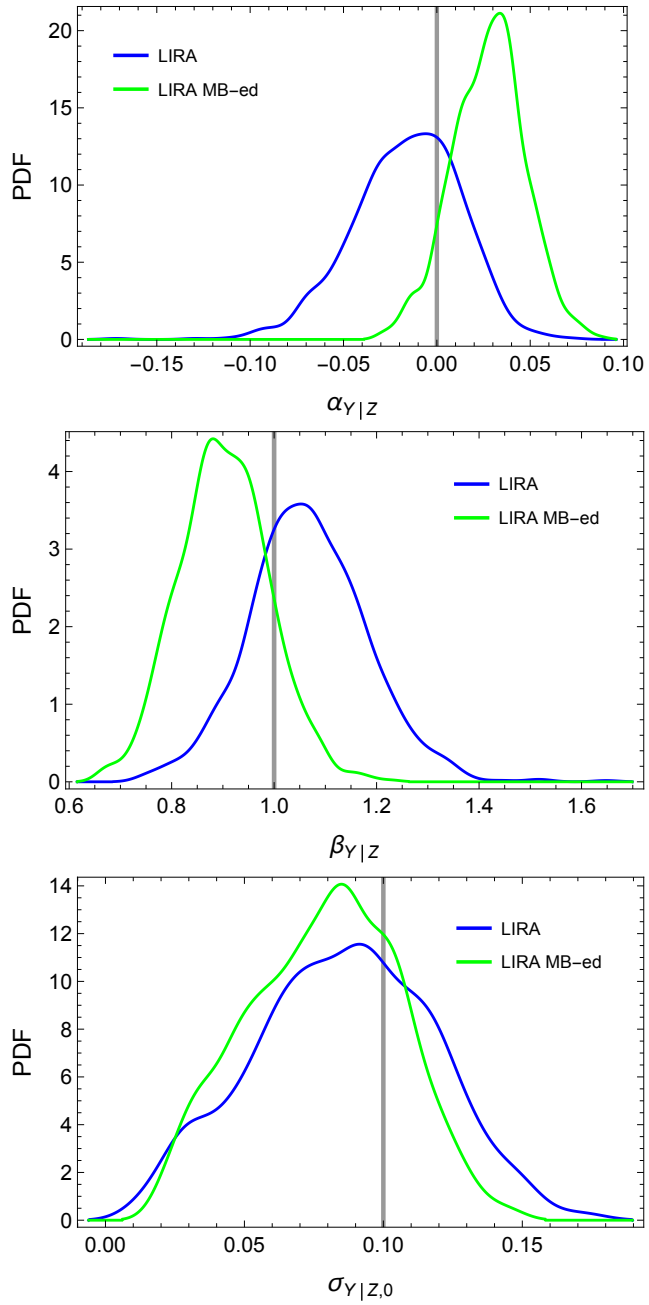

Figure 5. Distribution of the median parameters obtained from samples selected in the response variable. The blue line is the smoothed histogram of the distribution obtained by correcting for the Malmquist bias. The green line plots the results from a biased fit. The vertical gray lines are set at the values of the input parameters. From the top to the bottom panel: the intercept, the slope and the intrinsic scatter.

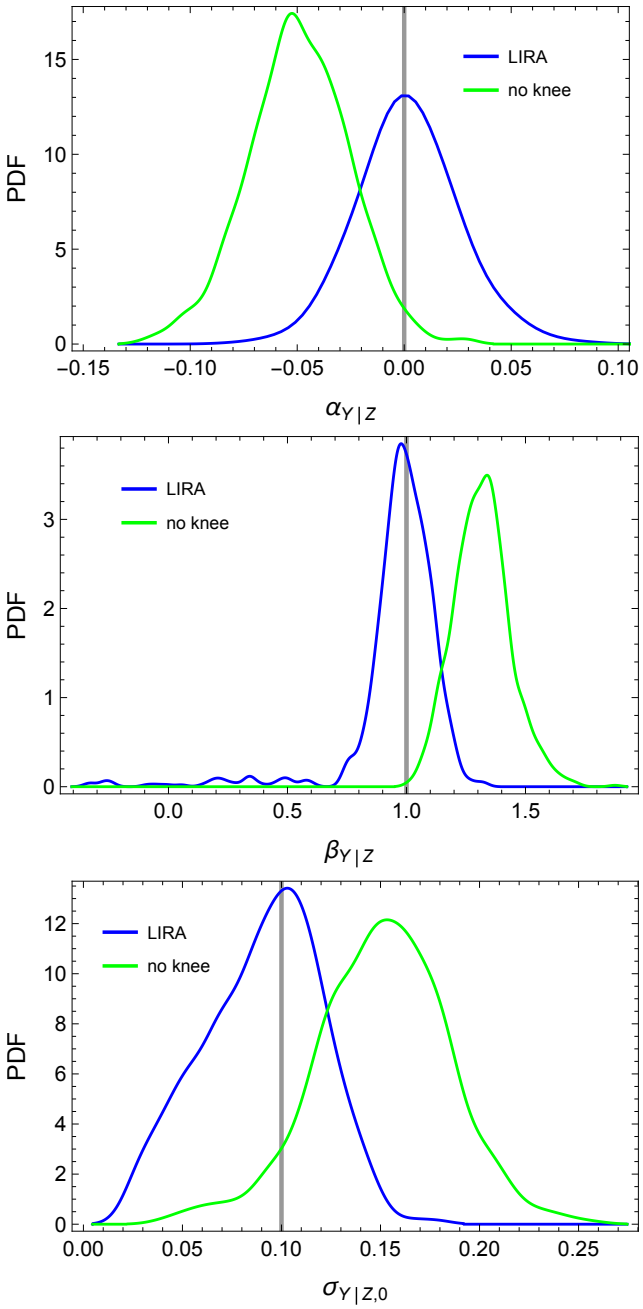

Figure 6. Distribution of the median parameters of a broken power-law. The blue line is the smoothed histogram of the distribution obtained by fitting the simulated data with a scattered broken power law. The green line plots the results from a biased linear fit. The vertical gray lines are set at the values of the input parameters. From the top to the bottom panel: the intercept, the slope and the intrinsic scatter.

4.5 Time dependent intrinsic scatter

Actual data sets are not rich enough to measure the time evolution of the intrinsic scatter (CoMaLit-IV). The γ parameters modeling the scatter redshift dependence, i.e., $\gamma_{\sigma_{Y|Z,F_z}}$ or $\gamma_{\sigma_{Y|Z,D}}$, are better seen as noise parameters to marginalize over.

The study of the time evolution of the scatter will be at reach of future surveys (Laureijs et al. 2011). I then increased the number of simulated sources per sample and their redshift range and I considered smaller observational errors. I simulated samples with 400 items each. The scale parameters of the scaled inverse χ^2-distributions modeling the uncertainty variances δx^2 and δy^2 were set to 0.05^2 and measurement errors were assumed to be uncorrelated.

Redshifts were drawn from a lognormal distribution such that $\ln z$ has mean $\ln(0.5)$ and standard deviation 0.8, i.e., $\sim$ 19 per cent of the sources are at $z > 1$. The independent variables were drawn from a time evolving normal distribution. The mean evolves

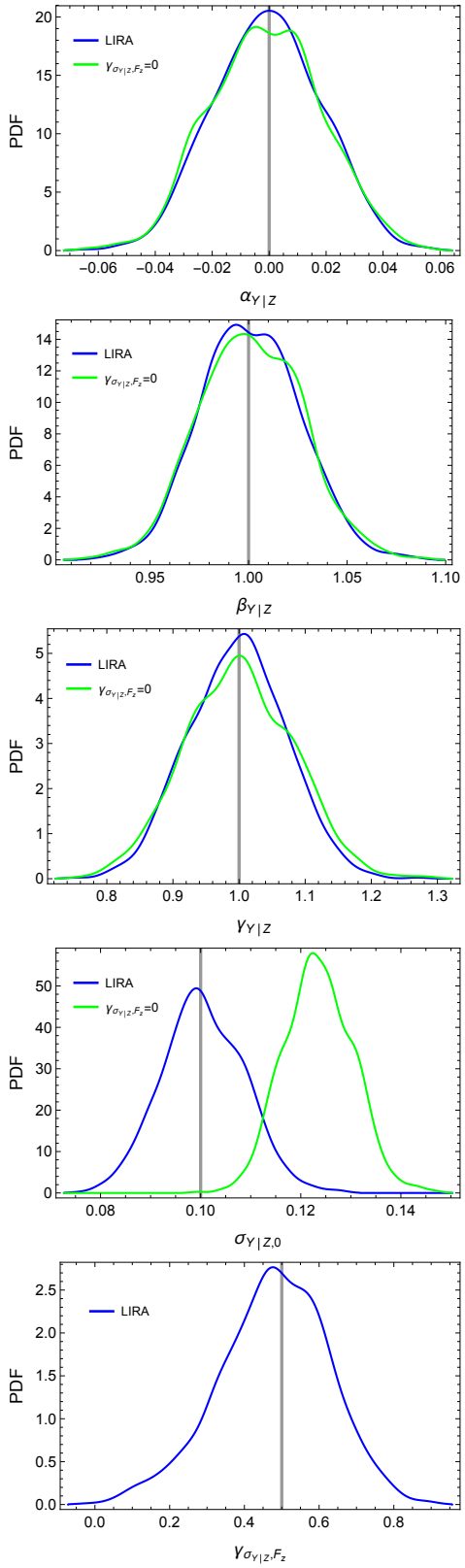

Figure 7. Distribution of the median parameters obtained from samples with time-evolving intrinsic scatter. The blue line is the smoothed histogram of the distribution obtained by fitting the simulated data with a time-dependent scatter. The green line plots the results from a biased linear fit with $\gamma_{\sigma_{Y|Z},F_z} = 0$. The vertical gray lines are set at the values of the input parameters. From the top to the bottom panel: the intercept, the slope, the time evolution, and the intrinsic scatter, and the scatter evolution.

Table 4. Scaling parameters recovered from time evolving samples. I report the bi-weight estimators of the distribution of the median values of the simulated chains.

parameter	input	unbiased	biased
	Time evolving scatter		
$\alpha_{Y\|Z}$	[0]	0.00 ± 0.02	0.00 ± 0.02
$\beta_{Y\|Z}$	[1]	1.00 ± 0.03	1.00 ± 0.03
$\gamma_{Y\|Z}$	[1]	1.00 ± 0.07	1.00 ± 0.08
$\sigma_{Y\|Z,0}$	[0.1]	0.100 ± 0.008	0.123 ± 0.007
$\gamma_{\sigma_{Y\|Z,F_z}}$	0.5	0.048 ± 0.15	[0]
	Time evolving slope		
$\alpha_{Y\|Z}$	[0]	0.00 ± 0.05	-0.05 ± 0.04
$\beta_{Y\|Z}$	[1]	1.00 ± 0.07	1.07 ± 0.05
$\gamma_{Y\|Z}$	[1]	0.93 ± 0.59	1.72 ± 0.34
$\delta_{Y\|Z}$	[1]	0.93 ± 0.67	[0]

with redshift as in Eq. (17) with $\gamma_{\mu_Z,F_z} = 0.5$ and $\gamma_{\mu_Z,D} = 0.5$; the standard deviation is constant.

The intrinsic scatter evolved with redshift as in Eq. (8), with $\sigma_{Y|Z,0} = 0.1$, $\gamma_{\sigma_{Y|Z,F_z}} = 0.5$ and $\gamma_{\sigma_{Y|Z,D}} = 0$. The input intrinsic scatter at $z \sim 1$ is ~ 30 per cent larger than the local value. The remaining parameters were set as for the other simulations.

Results are summarized in Table 4 and Fig. 7. Even if we neglect the scatter evolution, the estimates of the scaling parameters are unbiased whereas the estimated intrinsic scatter is weighted over the redshift range. The corrected regression can recover both the normalization and the time evolution of the scatter. Since the simulated sample is copious and observational accuracy is improved with respect to the other simulations, the posteriori distribution of the intrinsic scatter is symmetric, with no prominent tail at small values.

4.6 Redshift dependent slope

The emergence of some processes at high or low redshift might induce a tilting slope. I simulated a scaling relation with $\delta_{Y|Z} = 1$. The slope changes by $\Delta\beta_{Y|Z} \sim 0.25$ from redshift 0 to 1.

Redshifts were drawn from a lognormal distribution such that $\ln z$ has mean $\ln(0.3)$ and standard deviation 0.5. The independent variables were drawn from a time evolving normal distribution with $\gamma_{\mu_Z,F_z} = 0.5$ and $\gamma_{\mu_Z,D} = 0.5$; the standard deviation is constant. The scale parameters of the scaled inverse χ^2-distributions modeling the uncertainty variances δx^2 and δy^2 were set to 0.05^2 and measurement errors were assumed to be uncorrelated. The remaining parameters were set as in the basic scheme.

Results are summarized in Table 4 and Fig. 8. A correct modeling of the tilt is crucial to get unbiased parameters. Only the estimate of the intrinsic scatter is not affected.

5 CONCLUSIONS

Bayesian linear regression models involve a quite large number of parameters. The analysis of the hierarchical models can be performed with Markov Chain Monte Carlo (MCMC) simulations. Since all relations in the model are expressed as conditional probabilities, the posterior can be efficiently explored with a Gibbs sampler (Kelly 2007; Mantz 2015).

LIRA joins a number of already available routines for linear regression. Just to name a few of them which were proposed

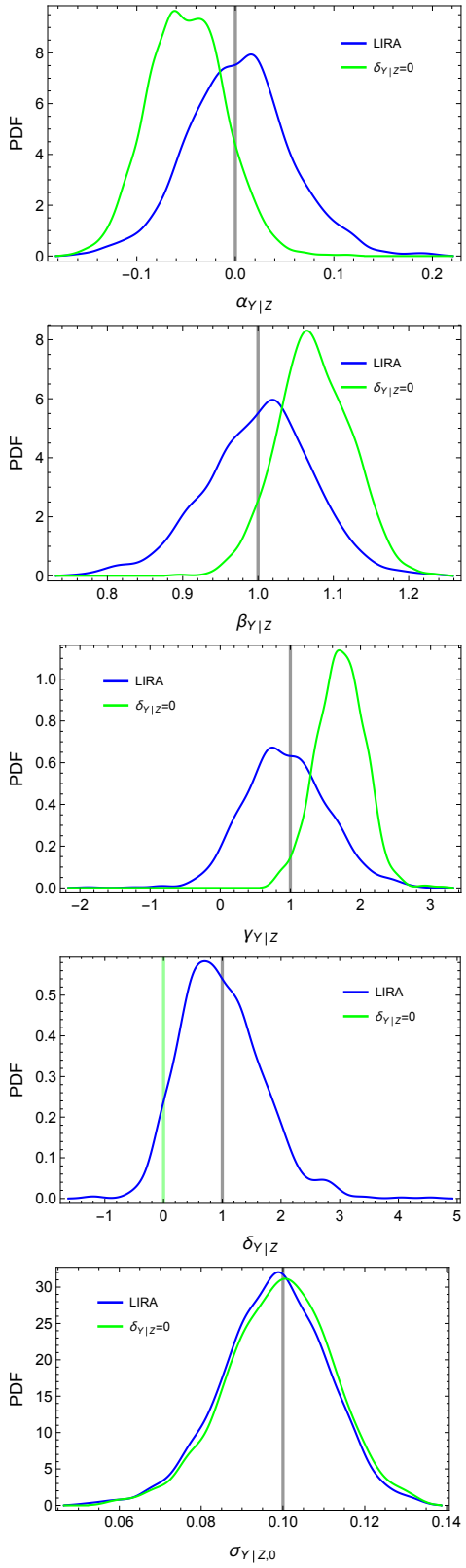

Figure 8. Distribution of the median parameters obtained from samples with time-evolving slope. The blue line is the smoothed histogram of the distribution obtained by fitting the simulated data with a time-dependent slope. The green line plots the results from a biased linear fit with $\delta_{Y|Z} = 0$. The vertical gray lines are set at the values of the input parameters. From the top to the bottom panel: the intercept, the slope, the time evolution, the slope tilt, and the intrinsic scatter.

to astronomers first, the Fortran function BCES, the IDL (Interactive Data Language) routine LINMIX and its multivariate extension MLINMIX, the Python package `astroML`[10](VanderPlas et al. 2012), and the R-packages LRGS[11], and HYPER-FIT[12].

All of these procedures have their own specifics and strengths that can make them preferable under given circumstances. LIRA is optimized for astronomical studies. It allows the consistent treatment of time-evolution, intrinsic scatter, and selection effects. Redshift has a prominent role in the proposed method. The time dependence of slopes, normalizations, intrinsic scatters, and correlations can be determined. Further complexity is implemented. The Malmquist and the Eddington biases can be addressed. Deviations from linearity and bent relations with knees can be accounted for.

The degree to which selection and methodological biases can affect the study of current and future samples was determined with a series of simulations. Selection effects are an important concern. But they are known problems and to some extent they are known unknowns. We usually know whether they are affecting our sample. Methodological bias can be unknown unknown. Different parameterization can give excellent fit to the data. Still results will differ and we do not know a priori the right parameterization. The problem is exacerbated by the high degree of degeneracy among involved parameters. The feature of a linear regression model to stay simple and to add complexity if needed is then important.

ACKNOWLEDGEMENTS

I thank S. Ettori for constant encouragement and stimulating discussions. M.S. acknowledges financial contributions from contracts ASI/INAF n.I/023/12/0 'Attività relative alla fase B2/C per la missione Euclid', PRIN MIUR 2010-2011 'The dark Universe and the cosmic evolution of baryons: from current surveys to Euclid', and PRIN INAF 2012 'The Universe in the box: multiscale simulations of cosmic structure'. S.E. acknowledges the financial contribution from contracts ASI-INAF I/009/10/0 and PRIN-INAF 2012 'A unique dataset to address the most compelling open questions about X-Ray Galaxy Clusters'.

REFERENCES

Akritas M. G., Bershady M. A., 1996, ApJ, 470, 706
Angulo R. E., Springel V., White S. D. M., Jenkins A., Baugh C. M., Frenk C. S., 2012, MNRAS, 426, 2046
de Souza R. S. et al., 2015, Astronomy and Computing, 12, 21
Eddington A. S., 1913, MNRAS, 73, 359
Eddington, Sir A. S., 1940, MNRAS, 100, 354
Evrard A. E., Arnault P., Huterer D., Farahi A., 2014, MNRAS, 441, 3562
Fabjan D., Borgani S., Rasia E., Bonafede A., Dolag K., Murante G., Tornatore L., 2011, MNRAS, 416, 801
Feigelson E. D., Babu G. J., 2012, Modern Statistical Methods for Astronomy. Cambridge Univ. Press, Cambridge
Giodini S., Lovisari L., Pointecouteau E., Ettori S., Reiprich T. H., Hoekstra H., 2013, Space Science Reviews, 177, 247
Hogg D. W., Bovy J., Lang D., 2010, arXiv:1008.4686
Isobe T., Feigelson E. D., Akritas M. G., Babu G. J., 1990, ApJ, 364, 104
Jeffreys H., 1938, MNRAS, 98, 190
Kelly B. C., 2007, ApJ, 665, 1489
Laureijs R. et al., 2011, arXiv:1110.3193

[10] `http://www.astroml.org/`
[11] `https://github.com/abmantz/lrgs`
[12] `https://github.com/asgr/hyper.fit`

Limousin M., Morandi A., Sereno M., Meneghetti M., Ettori S., Bartelmann M., Verdugo T., 2013, Space Science Reviews, 177, 155
Malmquist K. G., 1920, Lund Medd. Ser. II, 22, 1
Mantz A., Allen S. W., Rapetti D., Ebeling H., 2010, MNRAS, 406, 1759
Mantz A. B., 2015, arXiv:1509.00908
Mantz A. B. et al., 2015, MNRAS, 446, 2205
Maughan B. J., 2007, ApJ, 668, 772
Maughan B. J., 2014, MNRAS, 437, 1171
Pratt G. W., Croston J. H., Arnaud M., Böhringer H., 2009, A&A, 498, 361
Robotham A. S. G., Obreschkow D., 2015, arXiv:1508.02145
Rozo E., Bartlett J. G., Evrard A. E., Rykoff E. S., 2014, MNRAS, 438, 78
Saro A., Mohr J. J., Bazin G., Dolag K., 2013, ApJ, 772, 47
Sereno M., Ettori S., 2015a, MNRAS, 450, 3675
Sereno M., Ettori S., 2015b, MNRAS, 450, 3633
Sereno M., Ettori S., Moscardini L., 2015, MNRAS, 450, 3649
Sereno M., Ettori S., Umetsu K., Baldi A., 2013, MNRAS, 428, 2241
Stanek R., Rasia E., Evrard A. E., Pearce F., Gazzola L., 2010, ApJ, 715, 1508
VanderPlas J., Connolly A. J., Ivezic Z., Gray A., 2012, in Proceedings of Conference on Intelligent Data Understanding (CIDU), pp. 47-54, 2012., pp. 47–54
Vikhlinin A. et al., 2009, ApJ, 692, 1033

APPENDIX A: THE LIRA R-PACKAGE

The package LIRA is available through a GitHub[13] repository and can be installed from within R with the following command

```
> install_github("msereno/lira")
```

LIRA relies on the JAGS (Just Another Gibbs sampler) library[14], which must be installed separately, to perform the Gibbs sampling. C++ compilers are also needed.

The package is loaded into the R-session with

```
> library(lira)
```

The linear regression analysis is performed though the function `lira` (hence the name of the package), whose output are MCMC chains produced with a Gibbs sampling. Let `x`, `y`, `delta.x`, `delta.y`, `covariance.xy`, `z` be the vectors storing the values of $\boldsymbol{x}$, $\boldsymbol{y}$, δ_x, δ_y, δ_{xy} and $\boldsymbol{z}$, respectively. `n.data` is the length of the vectors.

- The chains analyzed in Sec. 4.1.1 and 4.1.2 were obtained with

```
> mcmc.lira <- lira(x, y, delta.x=delta.x,
delta.y=delta.y, covariance.xy=covariance.xy,
n.mixture=3)
```

and

```
> mcmc.lira <- lira(x, y, delta.x=delta.x,
delta.y=delta.y, covariance.xy=covariance.xy,
z=z, gamma.sigma.Z.Fz='dt'),
```

respectively.

- The case of Sec. 4.2 was studied with

```
> mcmc.lira <- lira(x, y, delta.x=delta.x,
delta.y=delta.y, covariance.xy=covariance.xy,
sigma.XIZ.0='prec.dgamma')
```

- The regression corrected for Malmquist bias, as in Sec. 4.3, is performed with

```
> mcmc.lira <- lira(x, y, delta.x=delta.x,
delta.y=delta.y, covariance.xy=covariance.xy,
y.threshold = rep(-0.3, n.data))
```

- The broken power-law in Sec. 4.4 was analyzed with

```
> mcmc.lira <- lira(x, y, delta.x=delta.x,
delta.y=delta.y, covariance.xy=covariance.xy,
Z.knee='dunif(-3.0,3.0)', beta.YIZ.knee ='dt')
```

- The samples with time dependent intrinsic scatters in Sec. 4.5 were analyzed with

```
> mcmc.lira <- lira(x, y, delta.x=delta.x,
delta.y=delta.y, z=z, gamma.sigma.XIZ.Fz='dt')
```

- The samples with time evolving slope in Sec. 4.6 were analyzed with

```
> mcmc.lira <- lira(x, y, delta.x=delta.x,
delta.y=delta.y, z=z, delta.YIZ='dt')
```

The LIRA package and further material and examples can also be found at `http://pico.bo.astro.it/~sereno/LIRA/`.

APPENDIX B: REDSHIFT EVOLUTION

In LIRA, the time-evolution of the parameters is factorized in two terms, one depending on F_z and one on the distance. The factor F_z can be either E_z or $(1+z)$. Since the redshift evolution is poorly constrained in present data-sets, and since both the cosmological distances and F_z are increasing function of the redshifts, the estimates of the evolution parameters γ_{F_z} and γ_D of each scatter/dispersion parameter are highly degenerate. It is usually enough to model just one dependence.

For limited redshift baselines, the function E_z can be approximated with a pawer law of $(1+z)$. The value of the exponent used in the approximation depends on the redshift range considered and on the cosmological parameters. It is very difficult to distinguish different features when modeling F_z as either E_z or $(1+z)$.

For similar reasons, the choice of the cosmological distance is secondary. The angular diameter and the luminosity distance differ for a factor $(1+z)^2$ which can be approximately englobed in $E_z^{\gamma_{F_z}}$ for limited redshift baselines.

[13] `https://hithub.com/msereno/lira`
[14] `http://mcmc-jags.sourceforge.net`.