Mon. Not. R. Astron. Soc. **000**, 1–7 (2013) Printed 13 August 2013 (MN LATEX style file v2.2)

Robust PCA and MIC statistics of baryons in early mini-haloes

R. S. de Souza[1], U. Maio[2,3], V. Biffi[4], B. Ciardi[5]
[1]*Korea Astronomy & Space Science Institute, Daedeokdae-ro 776, 305-348 Daejeon, Korea*
[2]*INAF - Osservatorio Astronomico di Trieste, Villa Bazzoni via G. B. Tiepolo 11, I-34143 Trieste, Italy*
[3]*Leibniz Institute for Astrophysics, An der Sternwarte 16, D-14482 Potsdam, Germany*
[4]*SISSA - Scuola Internazionale Superiore di Studi Avanzati, Via Bonomea 265, 34136 Trieste, Italy*
[5]*Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Str. 1, D-85748 Garching, Germany*

13 August 2013

ABSTRACT
We present a novel approach, based on robust principal components analysis (RPCA) and maximal information coefficient (MIC), to study the redshift dependence of halo baryonic properties. Our data is composed by a set of different physical quantities for primordial mini-haloes: dark-matter mass ($M_{\rm dm}$), gas mass ($M_{\rm gas}$), stellar mass ($M_{\rm star}$), molecular fraction ($\rm x_{mol}$), metallicity (*Z*), star formation rate (*SFR*) and temperature (T). We find that $M_{\rm dm}$ and $M_{\rm gas}$ are dominant factors for variance at high redshift, nonetheless, with the emergence of the first stars and subsequent feedback mechanisms, $\rm x_{mol}$, *SFR* and *Z* start to lead the variance. The PCA gives three principal components (PCs) that are capable to explain more than 97 per cent of the data variance at any redshift, while 2 PCs usually account for more than 92 per cent. Besides all the gaseous properties have a stronger correlation with $M_{\rm gas}$ than with $M_{\rm dm}$, our MIC analysis also suggests that $M_{\rm gas}$ has a deeper correlation with $\rm x_{mol}$ than with *Z* or *SFR*. This indicates the crucial role of gas molecular content to initiate star formation and consequent metal pollution from Pop III and Pop II/I regimes in primordial galaxies. Finally, a comparison between MIC and Spearman correlation coefficient shows that the former is a more reliable indicator when halo properties are weakly correlated.

Key words: theory: large-scale structure of Universe, early Universe; methods: statistical-cosmology

1 INTRODUCTION

The standard model of cosmology, predicts a hierarchal structure formation driven by cold dark matter (e.g., Benson 2010), where galaxies form from molecular gas cooling within growing dark-matter haloes. Hence, understanding the correlation between different properties of the dark-matter haloes is imperative to build up a comprehensive picture of galaxy evolution. Many authors have explored the correlation between dark-halo properties, such as mass, spin and shape, both in low- (e.g., Bett et al. 2007; Hahn et al. 2007; Macciò et al. 2007; Wang et al. 2011) and high-redshift (e.g., Jang-Condell & Hernquist 2001; de Souza et al. 2013a) regimes. Estimating the strength of these correlations is critical to support semi-analytical and halo occupation models, which assume the mass as determinant factor of the halo properties (e.g., Mo & White 1996; Cooray & Sheth 2002; Berlind et al. 2003; Somerville et al. 2008). Nevertheles, alternative approaches, based on principal components analysis (PCA), found that concentration is a key parameter, contrary to what expected before (Jeeson-Daniel et al. 2011; Skibba & Macciò 2011), and stressed the need for further investigations. PCA belongs to a family of techniques ideal to explore high-dimensional data. The method consists in projecting the data into a low-dimensional form, but keeping as much information as possible (e.g., Jollife 2002). Hence, PCA emerges as a natural technique to investigate correlation and temporal evolution of halo properties. Due to its versatility, PCA has been applied to a broad range of astronomical studies, such as stellar, galaxy and quasar spectra (e.g., Chen et al. 2009; McGurk et al. 2010), galaxy properties (Conselice 2006; Scarlata et al. 2007), Hubble parameter and cosmic star formation reconstruction (e.g., Ishida et al. 2011; Ishida & de Souza 2011), and supernova photometric classification (Ishida & de Souza 2013).

Notwithstanding PCA is not the only way to handle huge data sets, and the growth in complexity of scientific experimental data makes the ability to extract newsworthy and meaningful information an endeavor *per se*. The yearning for novel methodologies of data-intensive science gave rise to the so-called fourth research paradigm (e.g., Bell et al. 2009). Data-mining methods have been used in many areas of knowledge such as genetics (e.g., Venter et al. 2004) and financial marketing decisions (e.g., Shaw et al. 2001), and their importance for astronomy has been recently highlighted as well (e.g., Ball & Brunner 2010; Graham et al. 2013). Likewise observations, cosmological simulations are continuously increasing in complexity, lessening the distance between observed and synthetic data (e.g., Overzier et al. 2013; de Souza et al.

2013b). None the less, the application of data-mining to cosmological simulations remains a *terra incognita*.

In this Letter, we investigate the statistical properties of baryons inside high-z haloes, including detailed chemistry, gas physics and stellar feedback. We make use of Robust PCA (RPCA) and maximal information coefficient (MIC) to study a set of various halo parameters. While RPCA represents a generalization of the standard PCA, whose advantage is its resilience to outliers and skewed data, MIC is expected to be the correlation analysis of the 21st century (Speed 2011), in particular due to MIC ability in quantifying general associations between variables. Therefore, this project represents the first application of MIC to N-body/hydro simulations, and the first use of PCA to explore the low-mass end of the halo mass function and the birth of the first galaxies.

The outline of this paper is as follows. In Section 2, we describe the cosmological simulations. In Section 3, we introduce the statistical methods. In Section 4, we present our analysis and main results. Finally, in Section 5, we present our conclusions.

2 SIMULATIONS

We analyze the results of a cosmological N-body, hydro, chemistry simulation (Maio et al. 2010, 2011; Maio & Iannuzzi 2011), that was run by means of a modified version of the smoothed-particle hydrodynamics code GADGET2 (Springel 2005). The modifications include relevant chemical network to self-consistently follow the evolution of e^-, H, H^+, H^-, He, He^+, He^{++}, H_2, H_2^+, D, D^+, HD, HeH^+ (e.g., Yoshida et al. 2003; Maio et al. 2006, 2007, 2009), ultraviolet background radiation, metal pollution according to proper stellar yields (He, C, O, Si, Fe, Mg, S, etc.), lifetimes, and stellar population for Pop III and Pop II/I regimes (Tornatore et al. 2007), radiative gas cooling from molecular, resonant and fine-structure transitions (e.g. Maio et al. 2007, and references therein) and stellar feedback (Springel & Hernquist 2003). The transition from the Pop III to the Pop II/I regime is determined by the value of the gas metallicity (Z) compared to the critical value Z_{crit} (e.g., Omukai 2000; Bromm et al. 2001), assumed to be $10^{-4} Z_{\odot}$[1]. The cosmic field is sampled at redshift $z = 100$, with dark-matter and baryonic-matter species in the cosmological standard framework. We consider snapshots in the range $9 \lesssim z \lesssim 19$, within a cubic volume of comoving side 0.7 Mpc, and 2×320^3 particles per gas and dark-matter species corresponding to particle masses of $42\ \mathrm{M}_{\odot} h^{-1}$ and $275\ \mathrm{M}_{\odot} h^{-1}$, respectively. The identification of the simulated objects is done by applying a Friends of Friends (FoF) technique and substructures are identified by using a SubFind algorithm (Dolag et al. 2009), which discriminates among bound and non-bound particles. In order to avoid numerical artifacts, we select only those structures in which the gas content is resolved with at least 300 gas particles. This usually corresponds to selecting only objects with a total number of particles of at least $\sim 10^3$ (see more discussions in **?**, and references therein). The simulation outcomes investigated here consist of seven parameters: dark-matter mass ($M_{\rm dm}$), gas mass ($M_{\rm gas}$), stellar mass ($M_{\rm star}$), star formation rate (*SFR*), gas metallicity (Z), gas temperature (T), and gas molecular fraction $\mathrm{x}_{\rm mol}$. We refer the reader to our previous works, where more details and additional analyses about halo spin and shape distribution (de Souza et al. 2013a), feedback mechanisms (Maio et al. 2011; Petkova & Maio 2012; Maio et al. 2013), primordial streaming motions (Maio et al. 2011), non-standard cosmologies (Maio et al. 2006; Maio & Iannuzzi 2011; Maio 2011; de Souza et al. 2013c) high-z luminosity function (Salvaterra et al. 2013; Dayal et al. 2013), early gamma ray bursts- (Campisi et al. 2011; de Souza et al. 2011a, 2012; Maio et al. 2012) and supernovae-host properties (de Souza & Ishida 2010; de Souza et al. 2011b), Lyα emitters (Jeeson-Daniel et al. 2012) and DLA-system chemical content (Maio et al. 2013) are presented and discussed.

[1] Although uncertain (Bromm & Loeb 2003; Schneider et al. 2003, 2006), results are usually not very sensitive to the precise value adopted (Maio et al. 2010).

3 STATISTICAL ANALYSIS

Robust Principal Components Analysis. The ultimate goal of PCA is to reduce the dimensionality of a multivariate data[2], while explaining the data variance with as few principal components (PCs) as possible. PCA belongs to a class of Projection-Pursuit (PP) methods, whose aim is to detect structures in multidimensional data by projecting them into a lower-dimensional subspace (LDS). The LDS is selected by maximizing a projection index (PI), where PI represents an *interesting feature* in the data (trends, clusters, hypersurfaces, anomalies, etc.). The particular case where variance (S^2) is taken as a PI leads to the classical version of PCA[3].

Given n measurements $x_1, \cdots, x_n$, all of them column vectors of dimension p, the first PC is obtained by finding a unit vector $\mathbf{a}$ which maximizes the variance of the data projected on it:

$$\mathbf{a_1} = \underset{||\mathbf{a}||=1}{\arg\max}\ S^2(\mathbf{a}^t x_1, \cdots, \mathbf{a}^t x_n), \qquad (1)$$

where t is the transpose operation and $\mathbf{a_1}$ is the direction of the first PC[4]. Once we have computed the $(k-1)th$ PC, the direction of the kth component, for $1 < k \leqslant p$, is given by

$$\mathbf{a_k} = \underset{||\mathbf{a}||=1, \mathbf{a}\perp\mathbf{a}_1, \cdots, \mathbf{a}\perp\mathbf{a}_{k-1}}{\arg\max}\ S^2(\mathbf{a}^t x_1, \cdots, \mathbf{a}^t x_n), \qquad (2)$$

where the condition of each PC to be orthogonal to all previous ones, ensures a new uncorrelated basis. In spite of these attractive properties, PCA have some critical drawbacks as the sensitivity to outliers (e.g., Hampel et al. 2005) and inability to deal with missing data (e.g., Xu et al. 2010). In order to overcome this limitation several robust version were created based on the PP principle (e.g., Croux et al. 2007). Instead of taking the variance as a projection index in Eq. (1), a robust[5] measure of variance is taken. Two common measures of robust variance are: the median absolute deviation (*MAD*),

$$MAD(z_1, \cdots, z_n) = 1.48 \underset{j}{\mathrm{med}} |z_j - \underset{i}{\mathrm{med}} z_i|, \qquad (3)$$

[2] A set of measurements on each of two or more variables.

[3] The PCs are computed by diagonalization of the data covariance matrix (Σ^2), with the resulting eigenvectors corresponding to PCs and the resulting eigenvalues to the variance *explained* by the PCs.
The eigenvector corresponding to the largest eigenvalue gives the direction of greatest variance (PC1), the second largest eigenvalue gives the direction of the next highest variance (PC2), and so on. Since covariance matrices are symmetric positive semidefinite, the eigenbasis is orthonormal (spectral theorem).

[4] $\underset{x}{\arg\max}\ f(x)$ is the set of values of x for which the function $f(x)$ attains its largest value.

[5] Robust statistics commonly use inter-quantile range or median absolute deviation instead of mean and standard deviation, in order to be resistant against outliers.

and the first quartile of the pairwise differences between all data points (*Q*),

$$Q(z_1, \cdots, z_n) = 2.22 \left\{ |z_i - z_j|; 1 \leqslant i < j \leqslant n \right\}_{\binom{2}{n}/4}, \quad (4)$$

where $\{z_1, \cdots, z_n\}$ is a given univariate dataset and the square root of *MAD* or *Q* gives a robust variance[6]. Hereafter all calculations of the PCs are performed using the grid search base algorithm (Croux et al. 2007) with *MAD* as a variance estimator, but using *Q* has no influence on our results.

Maximal information coefficient. The maximal information-based nonparametric exploration statistics represent a novel family of techniques to identify and characterize general relationships in data sets (Reshef et al. 2011). They introduce MIC as a new measure of dependence between two-variables, which possesses two desired properties for data exploration: (i) generality, the ability to capture a broad range of associations and functional relationships[7]; (ii) equitability, the ability to give similar scores to equally noisy relationships of different types[8].
MIC measures the strength of general associations, based in on the mutual information[9] (MI) between two random variables [10], A and B:

$$\mathrm{MI}(A, B) = \sum_{a \in A} \sum_{b \in B} p(a, b) \log \left(\frac{p(a, b)}{p(a)p(b)} \right), \quad (5)$$

where $p(a)$ and $p(b)$ are the marginal probability distribution functions (PDF) of A and B and $p(a, b)$ is the joint PDF respectively.
Consider D a finite set of ordered pairs, $\{(a_i, b_i), i = 1, \ldots, n\}$, partitioned into a x-by-y grid of variable size, G, such that are x-bins spanning a and y-bins covering b respectively.
The PDF of a particular grid cell is proportional to the number of data points inside that cell. We can define a characteristic matrix $M(D)$ of a set D as

$$\mathrm{M(D)}_{x,y} = \frac{\max(\mathrm{MI})}{\log \min\{x, y\}}, \quad (6)$$

representing the highest normalized mutual informations of D. The MIC of a set D is then defined as

$$\mathrm{MIC(D)} = \max_{0 < xy < B(n)} \left\{ \mathrm{M(D)}_{x,y} \right\}, \quad (7)$$

representing the maximum value of M subject to $0 < xy < B(n)$, where the function $B(n) \equiv n^{0.6}$ was empirically determined by Reshef et al. 2011[11].

4 RESULTS

Hereafter we discuss the relations between halo properties and their relative importance. Our matrix is composed of $\approx$ 1500 haloes, spanning the redshift range $9 \lesssim z \lesssim 19$, each containing at least $\sim 10^3$ particles. Each row of the matrix represents a halo and each column represents one of the halo properties. RPCA probes the entire matrix at once[12]. On the other hand, MIC is a pair-variable comparison, therefore requiring $N(N-1)/2$ operations, with N being the number of halo properties. It is worth to highlight here that each approach has its own advantages and disadvantages. RPCA is suitable for high-dimensional data, when a pair comparison becomes unfeasible, however the method only searches for linear relationships. While MIC finds general associations in data structures, but may be impractical to deal with a large amount of parameters.

PCA. Figure 1 shows the contribution of the first three PCs to S^2, as a function of redshift. While 3 PCs account for more than 97 per cent of S^2 at any redshift, 2 PCs explain more than 92 per cent except at $z \simeq 14$, when the contribution drops to 85 per cent. The sharp variation of the PCs around $z \simeq 14 - 16$ acts as a smoking gun for a global cosmological event. Indeed, this is a direct consequence of first star formation episodes and the interplay between chemical and mechanical feedback from the first stars, that take place around $z \simeq 15 - 20$ (Maio et al. 2010, 2011; Maio & Iannuzzi 2011). As molecules are produced over time, they lead gas collapse, stellar formation and metal pollution with consequent back reaction on the thermal behavior of the surrounding gas (see e.g., Maio et al. 2011; Maio & Iannuzzi 2011; **?**). This redshift represents an epoch of fast and turbulent growth of the metal filling factor, from $\sim 10^{-18}$ at $z \simeq 15$ to $\approx 10^{-12}$ at $z \simeq 14$ (see Fig. 1 from Maio et al. 2011). At the beginning, only the gas at high densities is affected by metal enrichment, due to *SF* concentration in these regions. As *SF* and metal spreading proceed, the surrounding lower-density environments are affected as well. Supernova heats high-density gas within star-forming sites and, consequently, hot low-density gas is ejected from star-forming regions by supernova winds.
Figure 2 shows the relative contribution of each parameter to PC1 and PC2. At $z = 19$, $\mathrm{M_{dm}}$ and $\mathrm{M_{gas}}$ dominate PC1, followed by a small contribution of T. Nevertheless, as gas collapses into potential wells, the relative contribution from $\mathrm{M_{gas}}$ increases surpassing $\mathrm{M_{dm}}$ at $z \approx 15$. The dominant contribution of Z and $\mathrm{x_{mol}}$ to PC1 at $z \approx 14$ indicates a critical epoch for the cosmic chemical enrichment, triggered by a rapid variation of $\mathrm{x_{mol}}$, followed by a wide metal pollution at $z \approx 13$. After a decline in the chemical enrichment rate, a second peak in Z occurs at $z \approx 10$. This self-regulated, oscillatory behavior is caused by the simultaneous coexistence of

[6] When the PI is the standard variance, the first PC is the eigenvector of the data covariance matrix corresponding to the largest eigenvalue. But this does not hold for general choices of variance and approximative algorithms are necessary.
[7] For comparison, Pearson coefficient measures the linear correlation between two variables, while Spearman coefficient (R_s) measures the strength of monotonicity between paired data.
[8] In benchmark tests MIC equitability behaves better than other methods such as e.g., mutual information estimation, distance correlation and R_s. A lack of equitability introduces a strong bias and entire classes of relationships may be missed (Reshef et al. 2013).
[9] Mutual information measures the general dependence one variable contains about another, while the correlation function measures the linear dependence between them (e.g., Li 1990).
[10] MIC tends to 1 for all never-constant noiseless functional relationships and to 0 for statistically independent variables.
[11] The 0.6 exponent value represents a compromise since high values of$B(n)$ lead to non-zero scores even for random data, as each point gets its own cell, while low values only probes simple patterns.
[12] Before apply the RPCA, we standardize the halo properties by subtracting the means and dividing by the standard deviation. Therefore we are formally using the correlation matrix that can be seen as the covariance matrix of standardized variables.

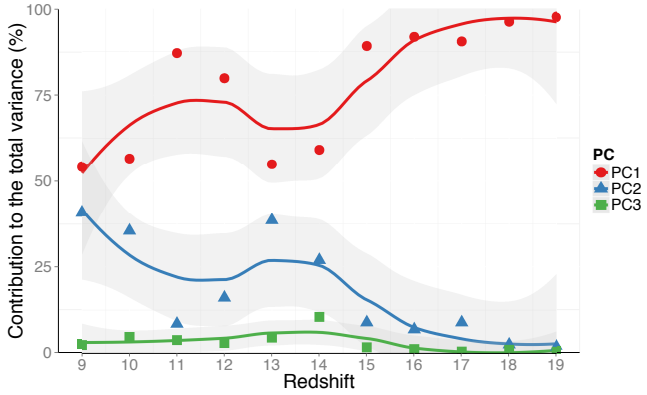

Figure 1. Fraction of variance explained by the first 3 PCs as a function of redshift. Symbols represent the actual estimate values for each snapshot, while the curves represent a smooth fitting with 95 per cent C.L. limited by the shadow areas.

cold pristine-gas inflows and hot metal enriched outflows that create hydro instabilities and turbulent patterns with Reynolds numbers $\sim 10^8 - 10^{10}$ (see e.g. Fig. 2 from Maio et al. 2011). Finally at $z = 9$, $\mathrm{M_{dm}}$ and $\mathrm{M_{gas}}$ have become almost subdominant, since PC1 is mainly led by T and Z, as a result of the ongoing cosmic heating from star formation and thermal feedback. An inspection in PC2 (right panel of Fig. 2) reveals the *supporting roles* during the galaxy formation process. The PC1 peak in Z at redshift 13 is preceded by a strong contribution of *SFR* and halo masses to PC2. While the second PC1 peak in Z, around $z \simeq 10$, is anticipated by an increasing contribution to PC2 by the formed stars, which later explode as supernovae and enrich the Universe.

MIC. Figures 3 and 4 show the correlation between the halo properties at $z = 10$ and 17 respectively[13]. The main diagonal shows the distribution of each variable[14], with the left vertical axis displaying the number of haloes per bin.

The lower triangular part of the panel shows scatter plots for each variable combination superimposed by density contours. This should facilitate a visual interpretation of the corresponding MIC and Spearman (R_s) coefficients quoted in upper triangular part of the panel. At high redshift, due to the poor statistics, most variables are uncorrelated, receiving a low score by both R_s and MIC. As expected $\mathrm{M_{gas}}$, $\mathrm{M_{dm}}$ and T are strongly correlated, receiving higher scores. Closely behind appear $\mathrm{x_{mol}}$, which is directly dependent on the local gas density and *T*, showing a moderate correlation with the 3 former quantities. An unexpected difference between the two approaches appears when comparing Z, $\mathrm{M_{star}}$ and the *SFR*. While R_s suggests a perfect correlation between Z and $\mathrm{M_{star}}$, MIC found no significant association at $z = 17$. This highlights the robustness of MIC with skewed and sparse data (in this redshift range, $z \gtrsim 17$, there are very few haloes with non-null Z and $\mathrm{M_{star}}$ values). Therefore, the high R_s value for these two quantities is misleading, as confirmed by a visual inspection of the corresponding distributions on the lower triangular part in Fig 3. During the course of cosmic evolution though, the correlations between the properties of the haloes tighten and both R_s and MIC converge for most of them at $z = 10$ (with R_s slightly overestimating the strength of correlation compared to MIC) as shown in Fig. 4.

[13] We do not display results for $z > 17$, since here there are too many zeros in the matrix and the variance measurements are unreliable.
[14] The variables are standardized by subtracting the mean, dividing by the standard deviation and transforming by $\log(1 + x)$ for better visualization.

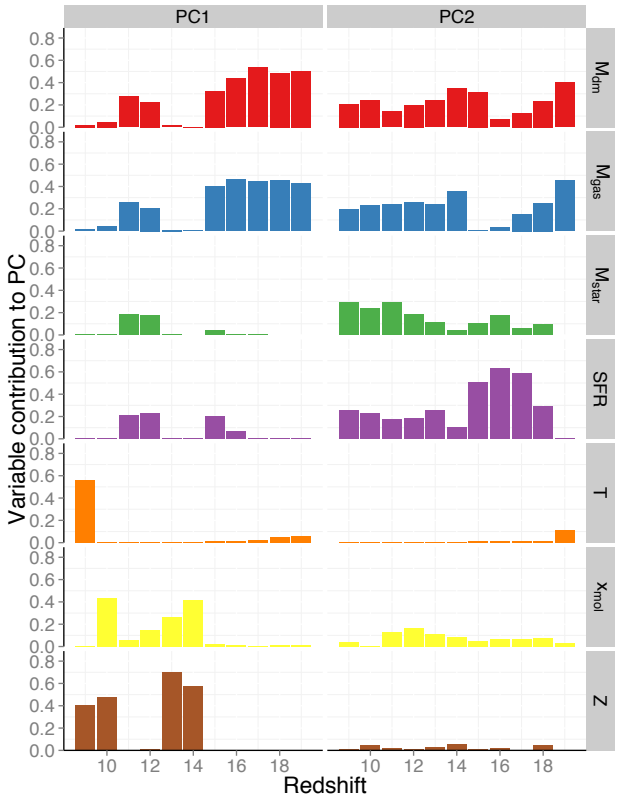

Figure 2. Variable contribution for PC1 and PC2 as a function of redshift.

5 CONCLUSIONS

We investigate the redshift evolution of the gas properties of primordial galaxies using robust PCA and MIC statistics.

This is the first attempt to probe the baryon properties of early mini-haloes and the effects of feedback processes by means of a statistically solid approach. We explore the correlation of different baryonic properties as expected from numerical N-body, hydrodynamical, chemistry simulations including gas molecular and atomic cooling, star formation, stellar evolution, metal spreading and feedback effects.

We find that two PCs are usually capable to explain more than 92 per cent of the data variance in the entire redshift range. The wide range of redshifts analyzed here ($9 \lesssim z \lesssim 19$) allowed us to study the temporal evolution of the relative contribution of each PC to the total variance. First star formation episodes and feedback mechanisms cause a drop of PC1 at $z \sim 14$, when a sharp variation in the PCs behavior marks the onset of cosmic metal enrichment. At $z > 14$ the halo properties are basically dictated by the halo mass.

Overall the Spearman correlation coefficient R_s agrees reasonably with MIC, but MIC seems to be more robust to study highly sparse data regimes (like at early epochs). All gas properties, aside $\mathrm{M_{gas}}$, $\mathrm{M_{dm}}$ and T, are weakly correlated at high redshift. Nevertheless, due to the interplay between chemical and mechanical feedback from the ongoing stellar formation and the consequent back reaction on the thermal behavior of the surrounding medium,

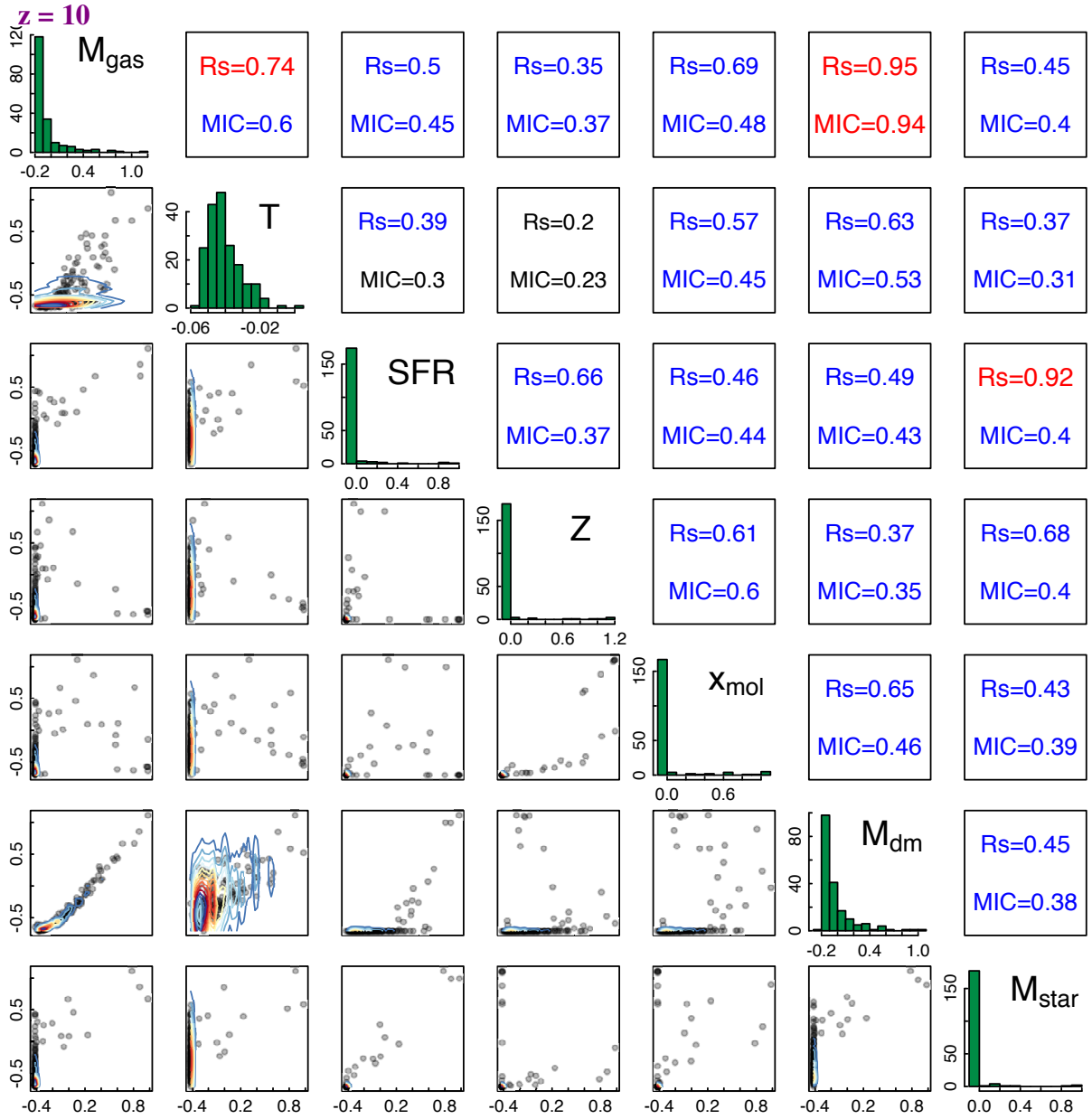

Figure 3. Correlations between different halo properties at redshift 10. The MIC and Spearman rank correlation coefficient are shown in the top half matrix. Values below 0.3 (weak correlation) are printed in black between 0.3-0.7 (moderate correlation) are printed blue, while values > 0.7 (strong correlation) are printed red. The panels on the diagonal show histograms of the parameter values. The bottom half matrix shows a scatter plot for each pair-variable combination. While the coefficients are estimate in the original parameters, the figures show the standardized variables transformed by $\log(1 + x)$ for better visualization.

baryonic quantities start to present a moderate to high level of correlation. In particular, $\mathrm{x_{mol}}$ shows the highest level of correlation with $\mathrm{M_{gas}}$, followed by T, *SFR*, M_{star} and Z respectively. In general, structure formation processes depend not only on the dark-matter halo properties, but also on the local thermodynamical state of the gas, which is, in turn, affected by cooling, star formation and feedback. Moreover, a combined inspection in the first and second PCs reveals some interesting facts. The PC1 peak in Z at redshift 13 is preceded by a strong contribution of *SFR* and halo masses to PC2. While the second PC1 peak in Z, around $z \simeq 10$, is anticipated by an increasing contribution to PC2 by the formed stars, which later explode as supernovae and enrich the Universe. Therefore stressing the importance of stellar evolution modeling in leading baryon properties in primordial haloes.

This work represents a leap forward in the statistical analysis of N-body/hydro simulation, performed by means of RPCA and MIC in a cosmological context. We therefore stress that the use of dimensionality reduction algorithms and mutual information based techniques in numerical simulations might be a precious instrument for future investigations, thanks to its potential to unveil non-trivial relationships, which may be inconspicuous to standard methods.

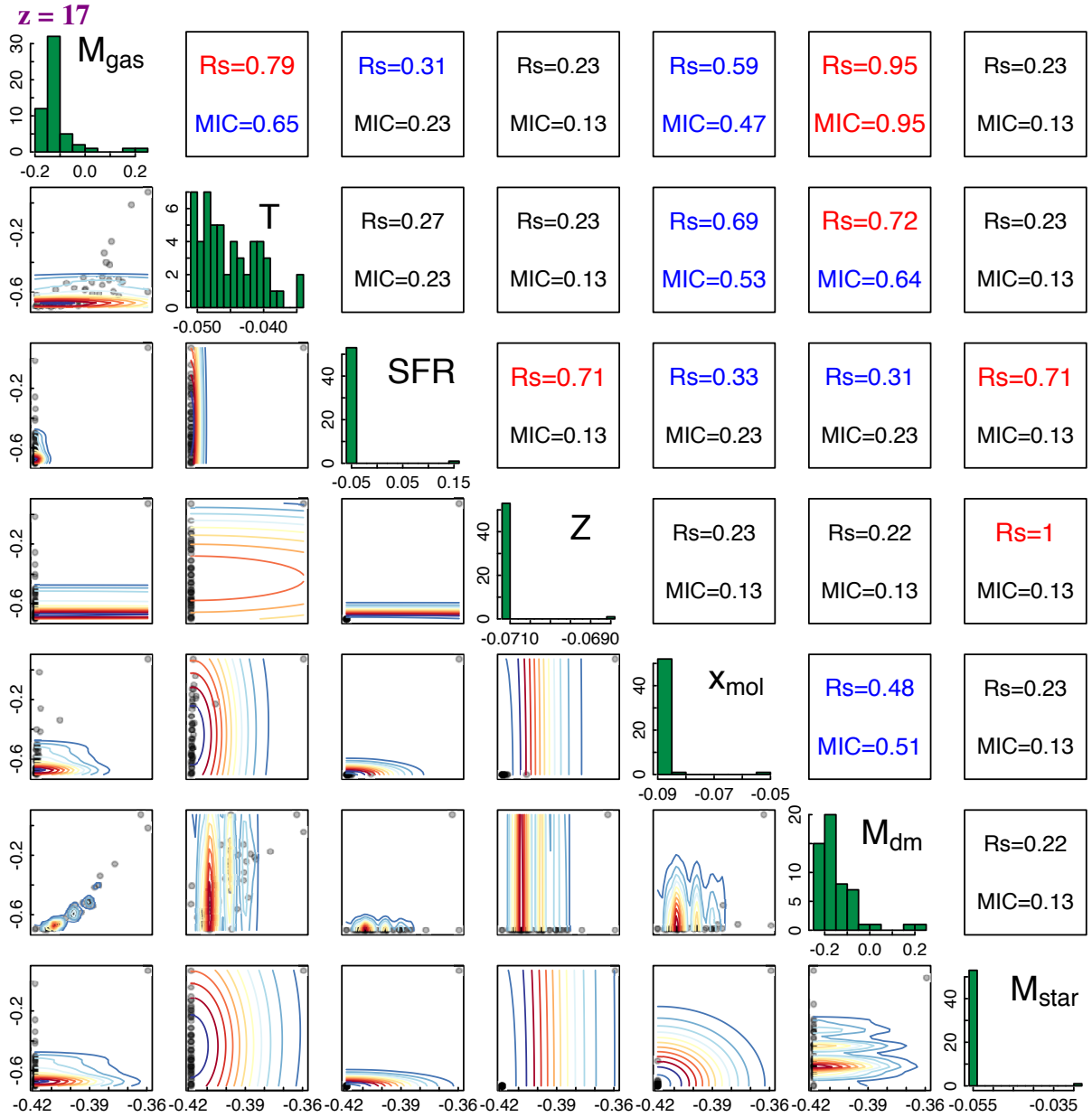

Figure 4. Same as in 3 at redshift 17.

ACKNOWLEDGEMENTS

We thank E.E. O. Ishida for revision of the manuscript. U.M. has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n.267251. For the biblyographic research we made use of the NASA Astrophysics Data System archive.

REFERENCES

Ball N. M., Brunner R. J., 2010, International Journal of Modern Physics D, 19, 1049
Bell G., Hey T., Szalay A., 2009, Science, 323, 1297
Benson A. J., 2010, Phys. Rep., 495, 33
Berlind A. A., Weinberg D. H., Benson A. J., Baugh C. M., Cole S., Davé R., Frenk C. S., Jenkins A., Katz N., Lacey C. G., 2003, ApJ, 593, 1
Bett P., Eke V., Frenk C. S., Jenkins A., Helly J., Navarro J., 2007, MNRAS, 376, 215
Bromm V., Ferrara A., Coppi P. S., Larson R. B., 2001, MNRAS, 328, 969
Bromm V., Loeb A., 2003, Nature, 425, 812
Campisi M. A., Maio U., Salvaterra R., Ciardi B., 2011, MNRAS, 416, 2760
Chen Y.-M., Wild V., Kauffmann G., Blaizot J., Davis M., Noeske K., Wang J.-M., Willmer C., 2009, MNRAS, 393, 406
Conselice C. J., 2006, MNRAS, 373, 1389
Cooray A., Sheth R., 2002, Phys. Rep., 372, 1
Croux C., Filzmoser P., Oliveira M., 2007, Chemometrics and Intelligent Laboratory Systems, 87, 218
Dayal P., Dunlop J. S., Maio U., Ciardi B., 2013, MNRAS
de Souza R. S., Ciardi B., Maio U., Ferrara A., 2013a, MNRAS, 428, 2109
de Souza R. S., Ishida E. E. O., 2010, A&A, 524, A74
de Souza R. S., Ishida E. E. O., Johnson J. L., Whalen D. J., Mesinger A., 2013b, arxiv:1306.4984

de Souza R. S., Krone-Martins A., Ishida E. E. O., Ciardi B., 2012, A&A, 545, A102
de Souza R. S., Mesinger A., Ferrara A., Haiman Z., Perna R., Yoshida N., 2013c, MNRAS, 432, 3218
de Souza R. S., Rodrigues L. F. S., Ishida E. E. O., Opher R., 2011b, MNRAS, 415, 2969
de Souza R. S., Yoshida N., Ioka K., 2011a, A&A, 533, A32
Dolag K., Borgani S., Murante G., Springel V., 2009, MNRAS, 399, 497
Graham M. J., Djorgovski S. G., Mahabal A. A., Donalek C., Drake A. J., 2013, MNRAS, 431, 2371
Hahn O., Porciani C., Carollo C. M., Dekel A., 2007, MNRAS, 375, 489
Hampel F. R., Ronchetti E. M., Rousseeuw P. J., Stahel W. A., 2005, Front Matter. John Wiley & Sons, Inc.
Ishida E. E. O., de Souza R. S., 2011, A&A, 527, A49
Ishida E. E. O., de Souza R. S., 2013, MNRAS, 430, 509
Ishida E. E. O., de Souza R. S., Ferrara A., 2011, MNRAS, 418, 500
Jang-Condell H., Hernquist L., 2001, The Astrophysical Journal, 548, 68
Jeeson-Daniel A., Ciardi B., Maio U., Pierleoni M., Dijkstra M., Maselli A., 2012, MNRAS, 424, 2193
Jeeson-Daniel A., Dalla Vecchia C., Haas M. R., Schaye J., 2011, MNRAS, 415, L69
Jollife I. T., 2002, Principal Component Analysis. Springer-Verlag, New York
Li W., 1990, Journal of Statistical Physics, 60, 823
Macciò A. V., Dutton A. A., van den Bosch F. C., Moore B., Potter D., Stadel J., 2007, MNRAS, 378, 55
Maio U., 2011, Classical and Quantum Gravity, 28, 225015
Maio U., Ciardi B., Dolag K., Tornatore L., Khochfar S., 2010, MNRAS, 407, 1003
Maio U., Ciardi B., Mueller V., 2013, arxiv:1307.6211
Maio U., Ciardi B., Yoshida N., Dolag K., Tornatore L., 2009, A&A, 503, 25
Maio U., Dolag K., Ciardi B., Tornatore L., 2007, MNRAS, 379, 963
Maio U., Dolag K., Meneghetti M., Moscardini L., Yoshida N., Baccigalupi C., Bartelmann M., Perrotta F., 2006, MNRAS, 373, 869
Maio U., Dotti M., Petkova M., Perego A., Volonteri M., 2013, ApJ, 767, 37
Maio U., Iannuzzi F., 2011, MNRAS, 415, 3021
Maio U., Khochfar S., Johnson J. L., Ciardi B., 2011, MNRAS, 414, 1145
Maio U., Koopmans L. V. E., Ciardi B., 2011, MNRAS, 412, L40
Maio U., Salvaterra R., Moscardini L., Ciardi B., 2012, MNRAS, 426, 2078
McGurk R. C., Kimball A. E., Ivezić Ž., 2010, AJ, 139, 1261
Mo H. J., White S. D. M., 1996, MNRAS, 282, 347
Omukai K., 2000, ApJ, 534, 809
Overzier R., Lemson G., Angulo R. E., Bertin E., Blaizot J., Henriques B. M. B., Marleau G.-D., White S. D. M., 2013, MNRAS, 428, 778
Petkova M., Maio U., 2012, MNRAS, 422, 3067
Reshef D., Reshef Y., Mitzenmacher M., Sabeti P., 2013, CoRR, abs/1301.6314
Reshef D. N., Reshef Y. A., Finucane H. K., Grossman S. R., McVean G., Turnbaugh P. J., Lander E. S., Mitzenmacher M., Sabeti P. C., 2011, Science, 334, 1518
Salvaterra R., Maio U., Ciardi B., Campisi M. A., 2013, MNRAS, 429, 2718
Scarlata C., Carollo C. M., Lilly S., Sargent M. T., Feldmann R., Kampczyk P., Porciani C., Koekemoer A., et al. 2007, ApJS, 172, 406
Schneider R., Ferrara A., Salvaterra R., Omukai K., Bromm V., 2003, Nature, 422, 869
Schneider R., Salvaterra R., Ferrara A., Ciardi B., 2006, MNRAS, 369, 825
Shaw M. J., Subramaniam C., Tan G. W., Welge M. E., 2001, Decision Support Systems, 31, 127
Skibba R. A., Macciò A. V., 2011, MNRAS, 416, 2388
Somerville R. S., Hopkins P. F., Cox T. J., Robertson B. E., Hernquist L., 2008, MNRAS, 391, 481
Speed T., 2011, Science, 334, 1502
Springel V., 2005, MNRAS, 364, 1105
Springel V., Hernquist L., 2003, MNRAS, 339, 289
Tornatore L., Borgani S., Dolag K., Matteucci F., 2007, MNRAS, 382, 1050
Venter J. C., Remington K., Heidelberg J. F., Halpern A. L., Rusch D., Eisen J. A., Wu D., Paulsen I., Nelson K. E., Nelson W., Fouts D. E., Levy S., Knap A. H., Lomas M. W., Nealson K., White O., Peterson J., Hoffman J., Parsons R., Baden-Tillson H., Pfannkoch C., Rogers Y.-H., Smith H. O., 2004, Science, 304, 66
Wang H., Mo H. J., Jing Y. P., Yang X., Wang Y., 2011, MNRAS, 413, 1973
Xu H., Caramanis C., Sanghavi S., 2010, arxiv:1010.4237
Yoshida N., Abel T., Hernquist L., Sugiyama N., 2003, ApJ, 592, 645

This paper has been typeset from a TEX/ LATEX file prepared by the author.