Abstract

Numerous upcoming missions will map our universe in unprecedented detail, and will create large and complex datasets with multiple dimensions. To efficiently and accurately analyze such datasets we develop a *group finding algorithm* which can work in a space of arbitrary number and type of dimensions. At the heart of all group finding schemes lies the choice of the distance metric. We develop a novel scheme which uses the information theoretic idea of *Shannon entropy* to calculate a *locally adaptive distance metric* for each data point which maximizes the information content extracted from the data. A density and nearest neighbor based scheme is then employed to identify groups in the data. As an application we apply this group finding algorithm to identify *tidal streams* produced by accretion of satellite galaxies in *simulated stellar halos* and in the 2MASS data. We also generate some synthetic surveys for upcoming missions like GAIA and show what we can learn about the stellar halo from such missions.

Group finding in multi-dimensional data-sets: method and application to Galactic Archeology

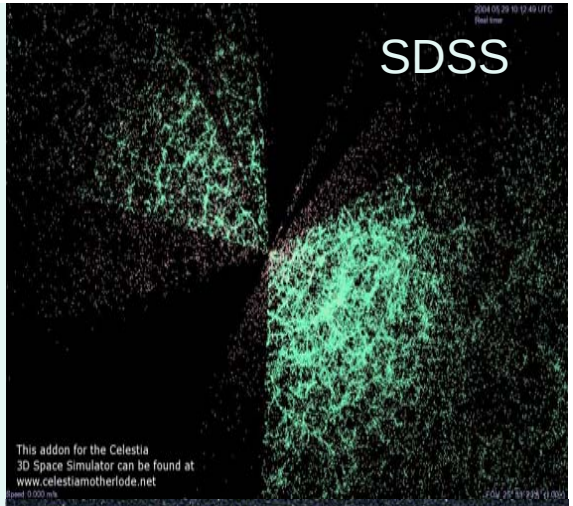
Sanjib Sharma

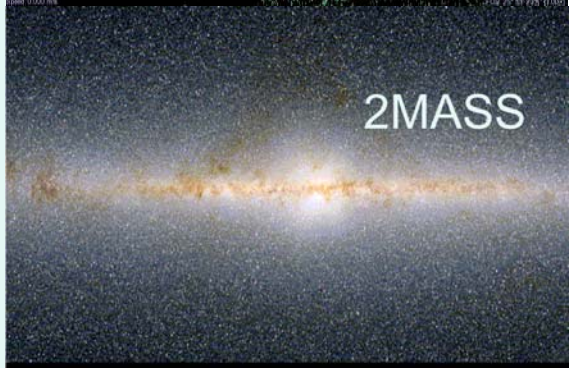
Kathryn V Johnston

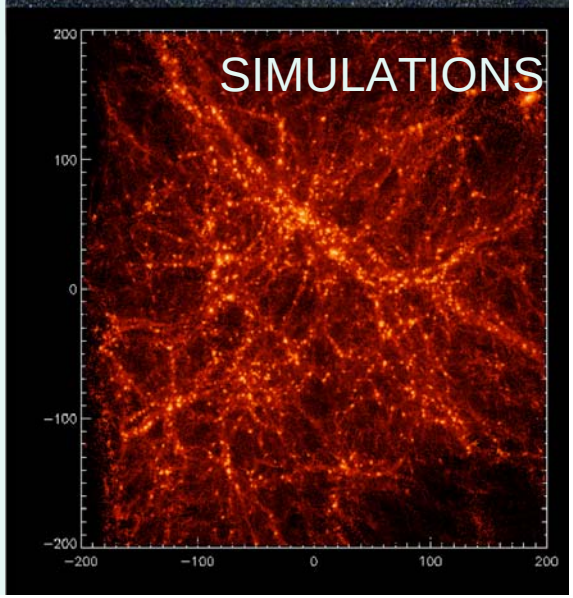
(Columbia University)

Motivation

- Due to advancement in technology, astronomy has become a data rich science. Numerous surveys with terabytes of information already exist and upcoming missions will create even larger and complex datasets (2MASS,SDSS, GAIA, SIM, WIFMOS). Moreover N-body simulations of structure formation at ever high resolution are also producing large amounts of data.
- Data sets are also increasing in complexity with data having arbitrary number and type of dimensions. For example there could be a catalogue of objects containing position co-ordinates ,velocity co-ordinates, radial velocity, angular momentum , chemical abundance, magnitude, color.
- There is a need for efficient methods and algorithms to analyze such data sets.
- *We describe here a group finding algorithm to identify clusters in such data sets. The algorithm is*
  - computationally fast so as to handle huge amounts of data
  - efficient at handling complex multiple dimensions
  - and can be easily parallelized to make use of distributed computing





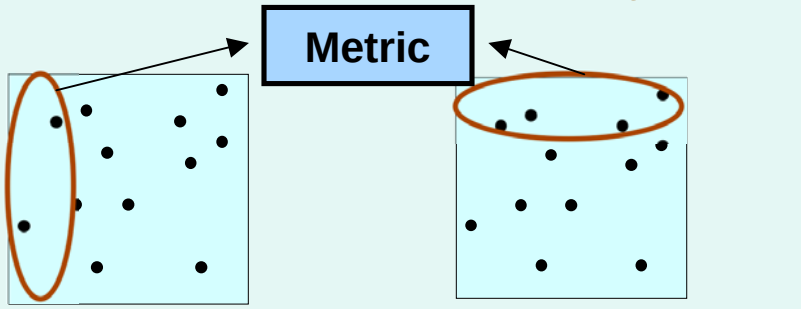


Why do we need a new group finding algorithm?

- All group finding schemes need an *a priori* definition of a metric. A metric is a non-negative function g_{ij} which gives the distance between any two neighboring points in a space.

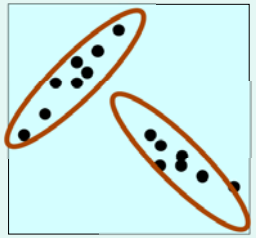
$$ds^2 = \sum_i \sum_j g_{ij} dx_i dx_j = \mathbf{dx}^{\mathrm{T}} \mathbf{G}\, \mathbf{dx}$$

- Improper choice of metric can wash out structures or lead to poor quality of structures.



x information lost y information lost

- Global scaling not suitable for
  - all data points in space. Strictly correct only for the mean.
  - systems with non-gaussian distributions.
  - systems with multiple anisotropic structures.



Multiple anisotropic structures

- Curse of dimensionality
  - The spatial resolution per data point reduces exponentially with the number of dimensions *d*. For a cube of length *L* containing *N* particles, mean interparticle separation $l = L / N^{1/d}$. For $N=10^6$, $l=0.1, 0.001$ and 10^{-6} for $d=1,3$ and 6 respectively.
  - Incorrect choice of metric exponentially erodes the resolution.
  - Uninformative hidden dimensions need to removed. Global dimensionality reduction techniques again not suitable for multi-component systems.

How to calculate the optimum locally adaptive metric?

- If $\sum(x)$ is the local covariance matrix at a point x then the ideal metric is given by $\sum^{-1}(x)$ (Mahalanobis metric).

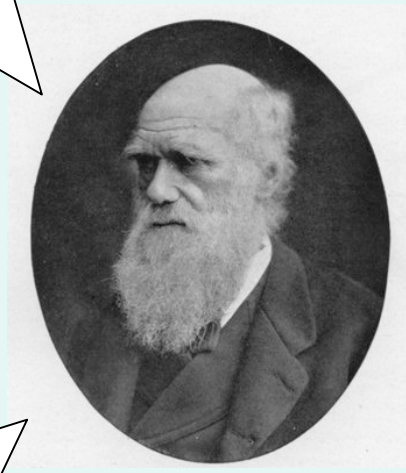
$$ds^2 = dx^T \sum^{-1}(x)\, dx$$

$$g_{\alpha\beta}(x) = \sum^{-1}{}_{\alpha\beta}(x)$$

- But to calculate the local neighborhood itself one needs a definition of a metric.

- We develop a *heuristic algorithm* which calculates the local neighborhood or the local $\sum$ without an *a-priori* definition of a metric.

Which comes first, the *chicken* $\sum$ or the *egg* $g_{\alpha\beta}$?



An evolutionary theory (algorithm) that generates *chicken* $\sum$ which then lays the *egg* $g_{\alpha\beta}$.

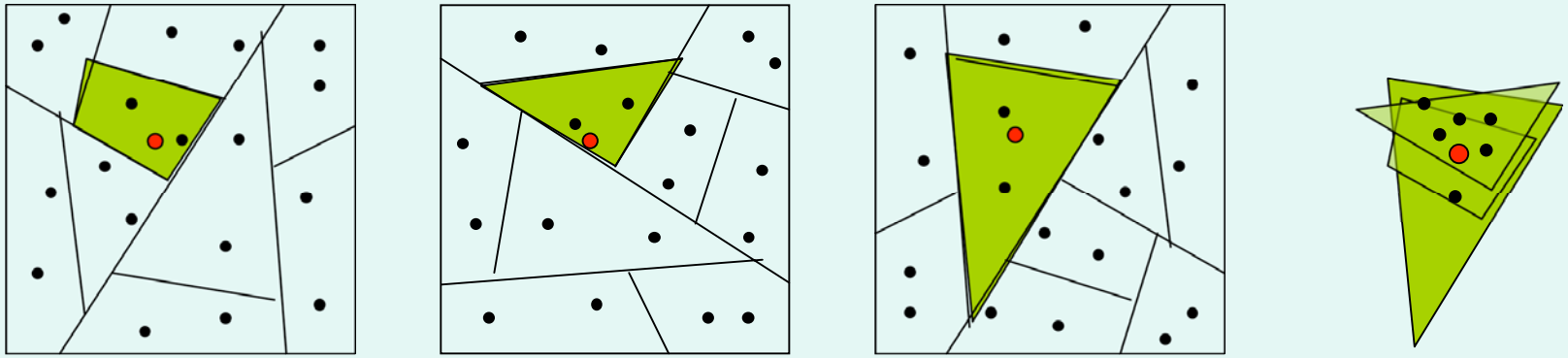
Algorithm for calculating the optimum metric

- The algorithm EnBiD (Entropy based Binary Decompostion)
  - Calculate the *covariance matrix* of the data points and orient the co-ordinate system along its eigenvectors.
  - Do an "Independent Component Analysis" ICA and orient the co-ordinate system along the independent axes. .
  - Calculate the *Shannon Entropy* (a measure of information content) along each dimension and identify the dimension having minimum entropy in other words maximum amount of information.
  - Calculate the *mean co-ordinate* of the data points along this dimension.
  - Split the data space into two equal parts by means of a *hyper-plane* perpendicular to the above dimension and passing through this point.
  - *Recursively* repeat the above procedure on the sub-partitions till each partition contains some predefined number of data point.

NOTE: Orienting the co-ordinates automatically takes care of the anisotropic structures and the Entropy criteria helps to scale out the uninformative dimensions.

For an older simpler version of the algorithm see ***Sharma & Steinmetz 2005*** and also see ***Ascasibar & Binney 2005*** where the idea of tree partitioning was first used.

- This generates a *binary space partitioning (BSP) tree* with each leaf node containing a fixed number of minimum data points.



Binary partitions of data with each leaf node containing 2 points. 3 random instances are shown. The colored area is the local neighborhood of the red point. Combining the three random instances one gets the average neighborhood shown.

- The local neighborhood of a given point x is the data points in its corresponding leaf node. Use these data points to calculate the local $\sum(x)$ and hence $g_{\alpha\beta}(x)$.
  - Metric thus obtained will be noisy and discontinuous.
- Use *bootstrapping* to obtain a smooth metric.
  - Calculate the covariance matrix $\sum$ several times by randomly sub-sampling the data set and then average it.

Group finding general outline

- Input:
  - a set of N points $X=\{x_0, x_1, \ldots x_N\}$ in a d dimensional space.
- Method: For each data point x
  - Calculate the appropriate metric $g_{\alpha\beta}(x)$.
  - Create a nearest neighbor list $l=\{l_0, l_1 \ldots l_k\}$.
  - Calculate the density $\rho(x)$.
  - Use a density based scheme along with nearest neighbor links to cluster the data points.
- Output:
  - Cluster labels for the N data points $\{c_0, c_1 \ldots c_i, \ldots c_N\}$ where $1 \le c \le C$, C being total number of clusters.

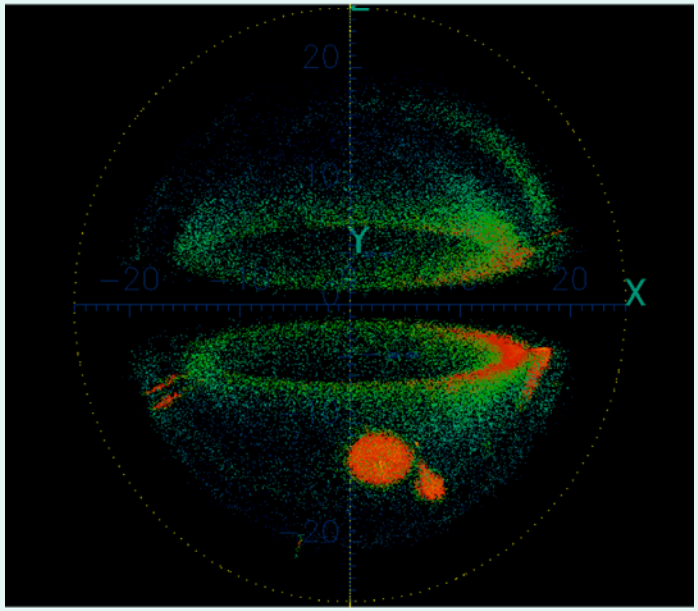
Application: Identifying substructures in the stellar halo of a galaxy

According to the hierarchical paradigm of structure formation, the stellar halo of our galaxy is thought to have been produced by trails of stars disrupted from satellite galaxies. For older systems strong phase mixing takes place which erodes the signatures of the accretion process in the configuration space. In phase space or integrals of motion space such structures are preserved for longer timescales hence it is easier to identify them. We first apply the group finder the 2MASS data and identify substructures in it. Next we generate some synthetic surveys from N-body simulations of the stellar halo (*Bullock & Johnston 2005*) and apply the group finder on it. We show results for spaces of different dimensions namely x y z space ,x y z v_r space , x y z v_r FeH α/Fe space and x y z v_x v_y v_z space. v_r here is the radial velocity , FeH the metalicity and α/Fe the abundance of α elements.

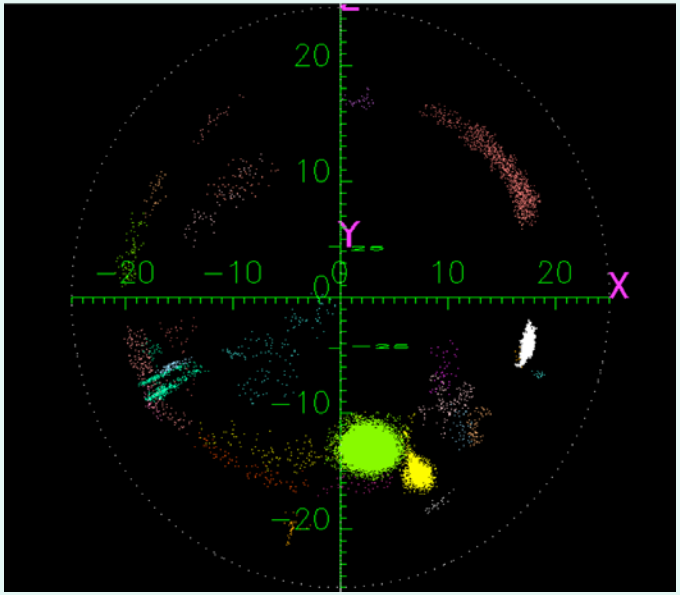
Substructures in 2MASS

Data: ***Majewski et al***

- 2MASS is an all sky uniform survey with photometry in three infrared bands *J* (1.25 microns), *H* (1.65 microns), and *Ks* (2.17 microns). It has a point source catalogue of 300 million stars. A color magnitude cut of $0.97 < J-K < 1.2$ and $K>10$ was applied which generated a sample of 60,000 stars.
- Distance was calculated assuming [Fe/H]=-1.
- Group finder was applied in 3d space with radial distance being measured in distance modulus *d=5log(r/10pc).*



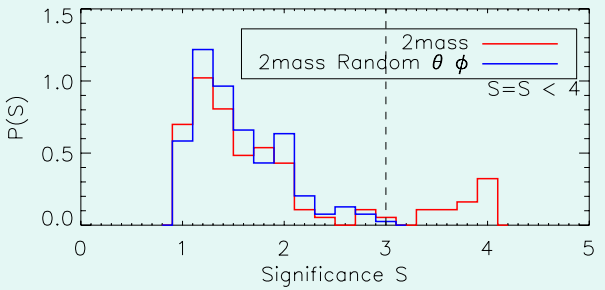
Stars color coded with 3d density

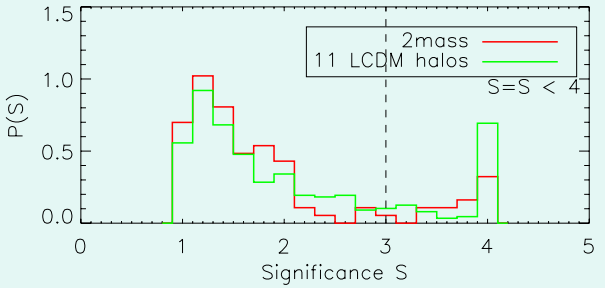


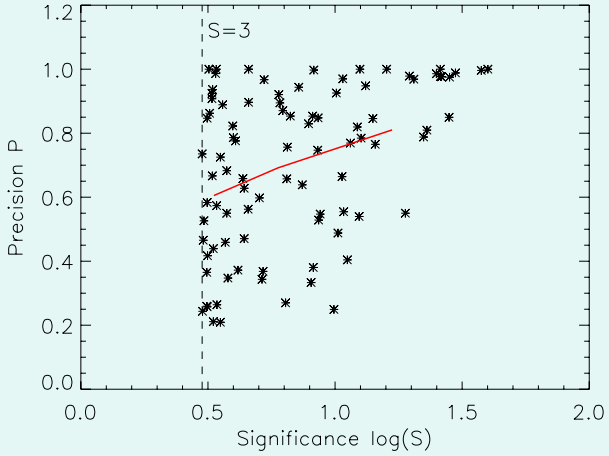
Groups found by group finder

2MASS group finder analysis

- The group finder gives a measure of significance of a group as $S=[\log(\rho_{max})-\log(\rho_{min})]/\sigma_{\log(\rho)}$. ρ_{max} and ρ_{min} being the max and minimum density of the particles in the group. Groups with S<3 are generally due to poisson noise.
- A histogram of S for a data with stars having random orientation shows that for S>3 the probability of a group being spurious is very low.
- Next we run the group finder on 2MASS data set and identify groups with S>3. We get 15 groups out of which 10 can be associated with known structures in the Milky Way while 3-5 groups are new structures. Simulated LCDM halos when subjected to similar analysis give on average 8 groups with standard deviation of about 4.
- We define Precision P= fraction of stars in a group which belong to a single satellite in the simulations. A plot of P Vs S shows that the precision of the group increases with S and for S>3 it is greater than 0.67.

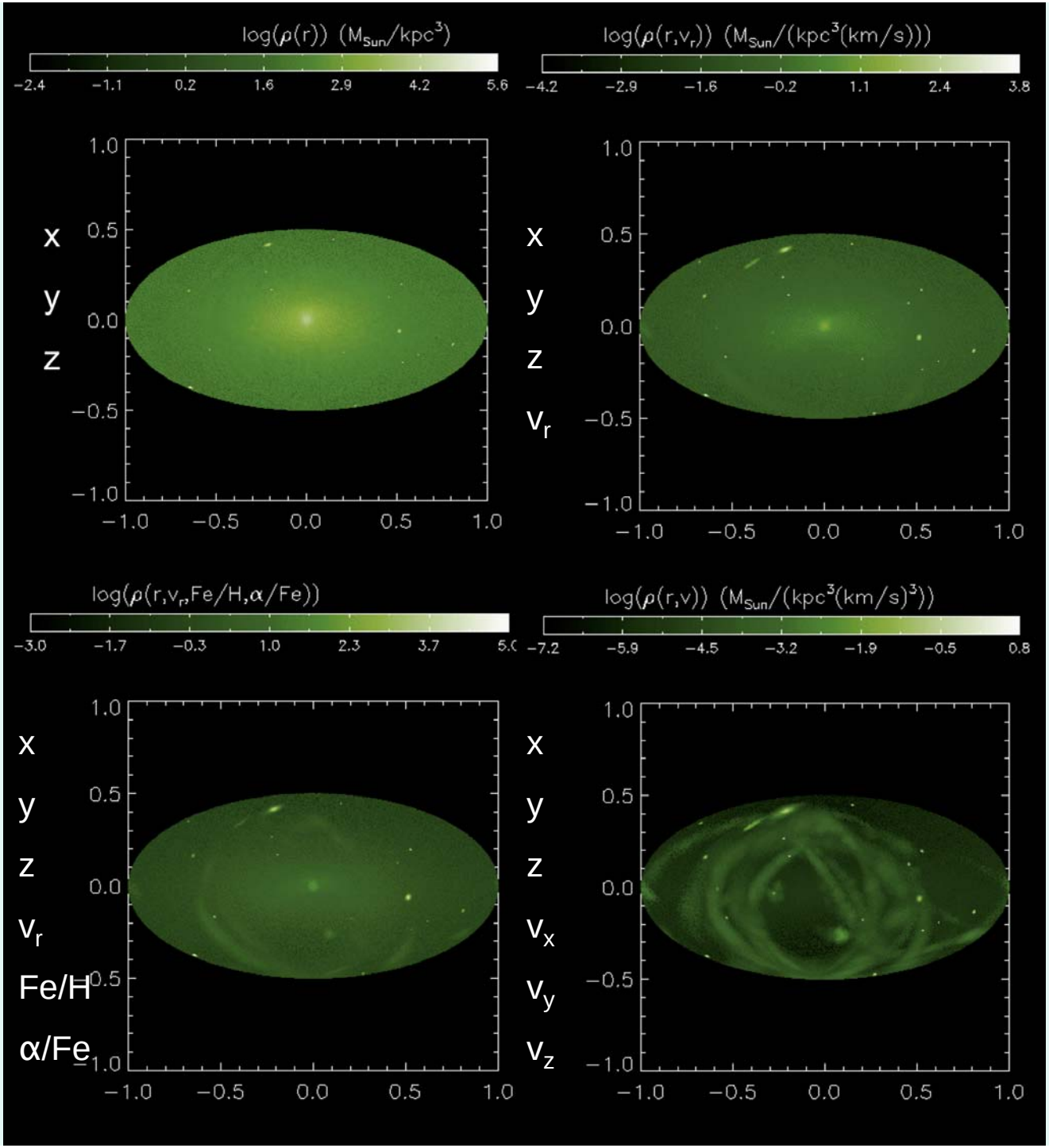




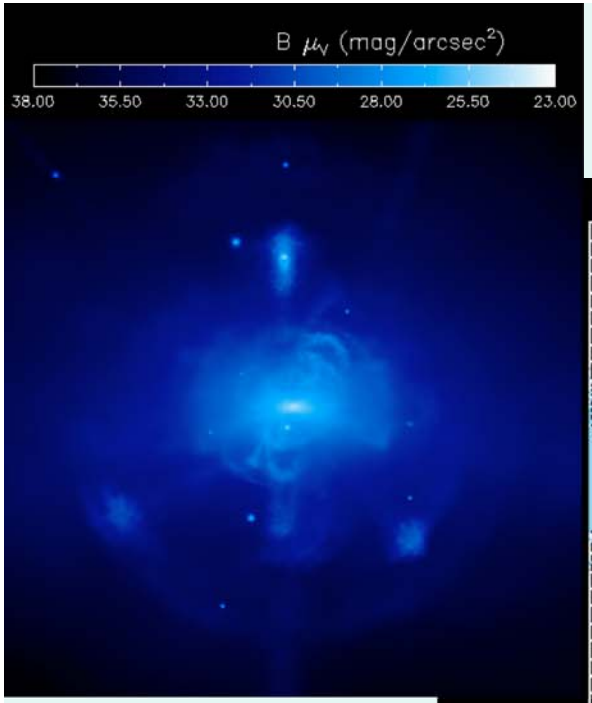


Effect of Dimensions

Aitoff projection of particles obtained from simulations of the stellar halo. Particles are color coded with density in various spaces. The densest particle being on the top. As the number of dimension increases the substructures are resolved better.



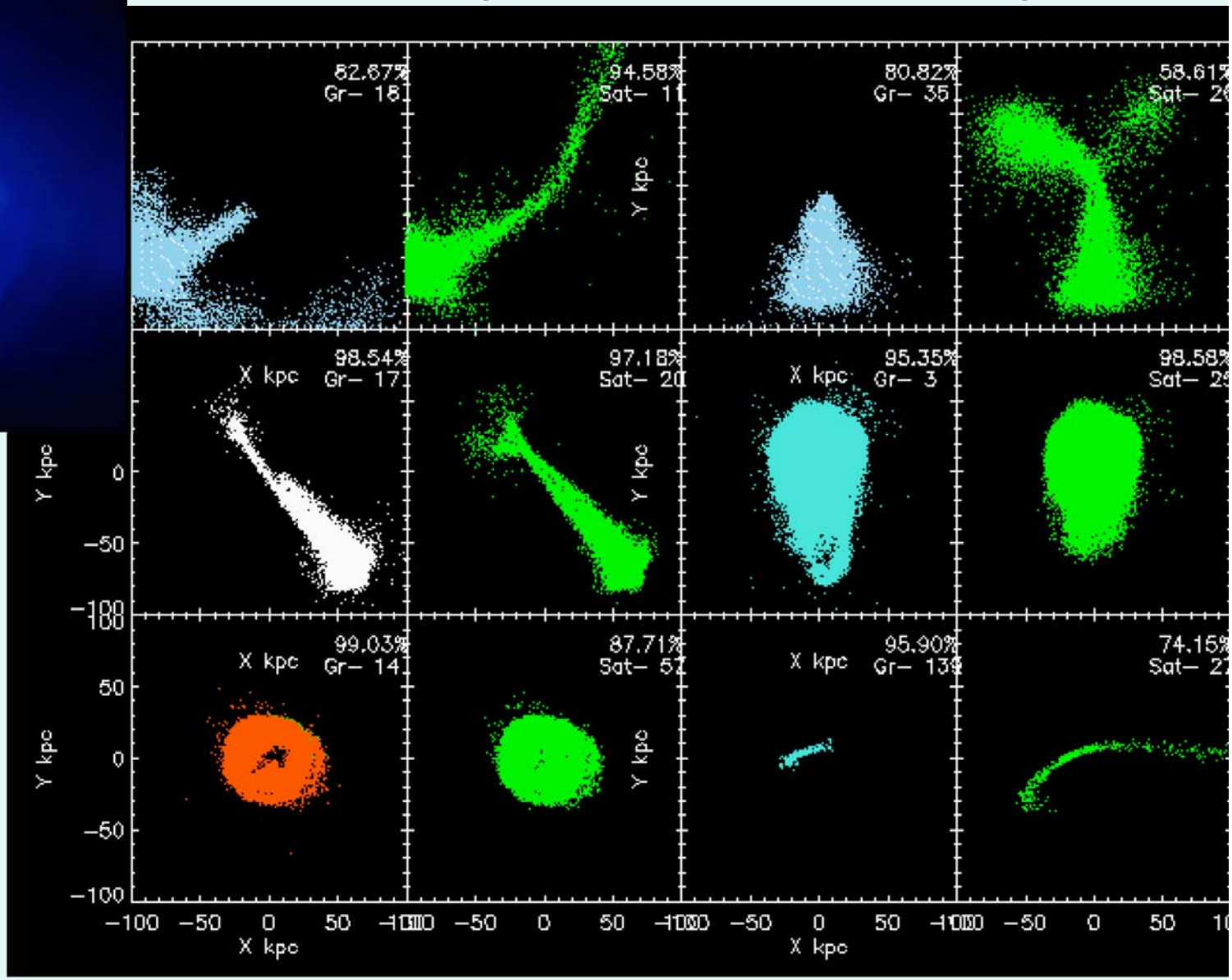
Substructures in a simulated stellar halo



3d density map

In spite of significant mixing in the **configuration space** the group finder can recover the **satellites**.

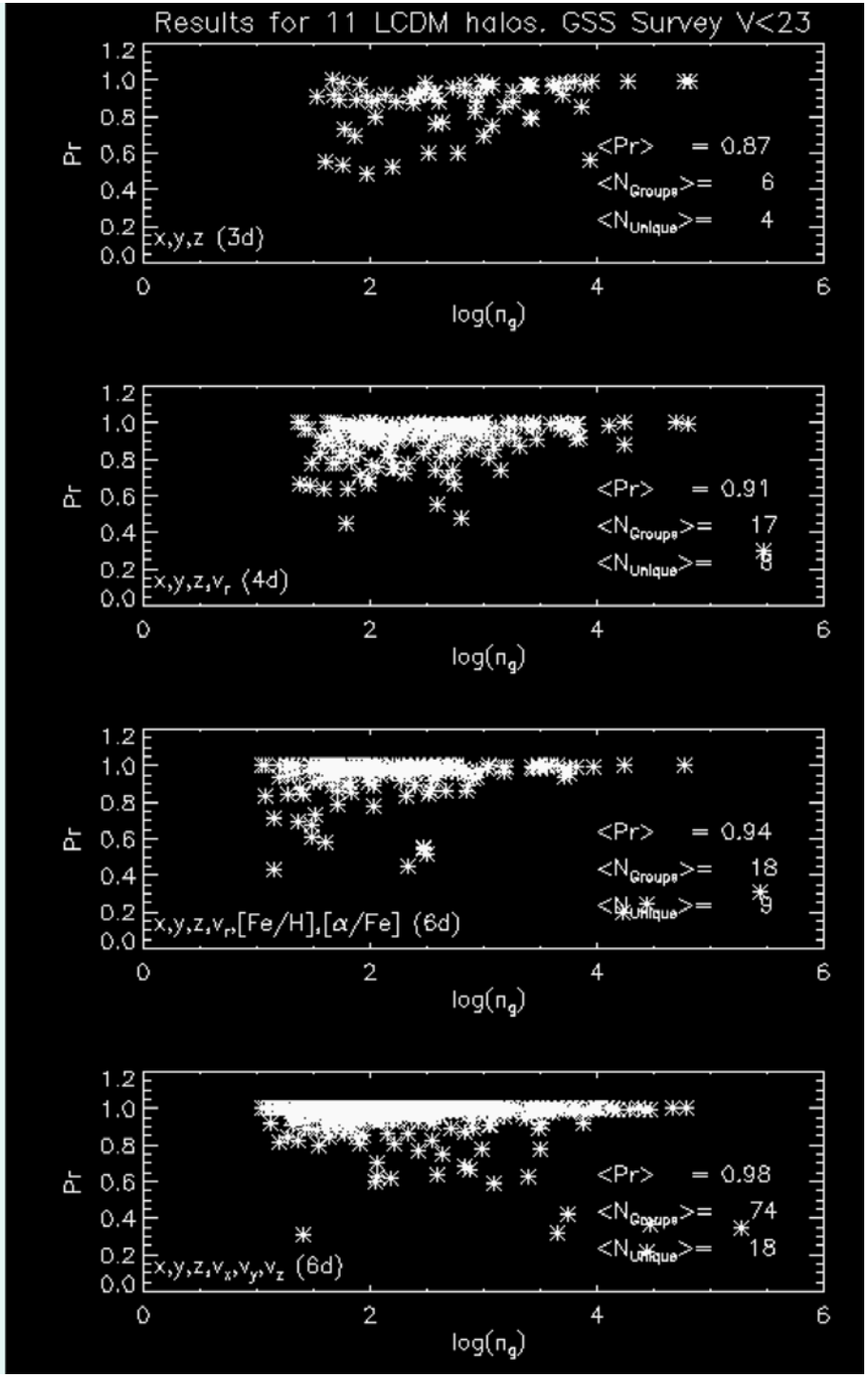
Recovered Original Recovered Original



The XZ Scatter plot of satellites identified by the group finder in 6d phase space.

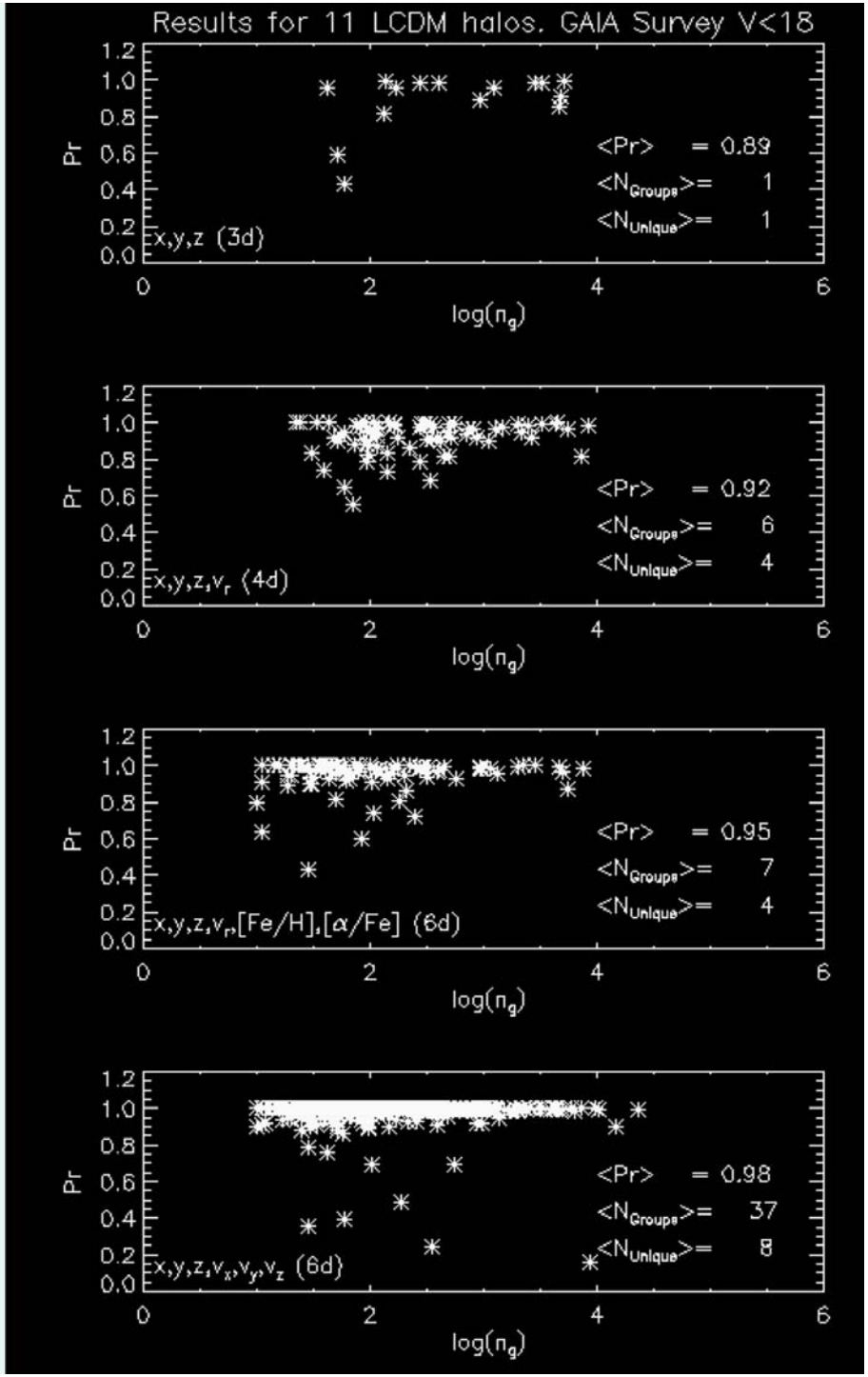
All Sky Giant Star Survey

- A set of 11 simulated stellar halos (Bullock & Johnston 2005) formed by accretion of satellite galaxies having merger history consistent with the LCDM paradigm were used to generate a synthetic survey of stars having absolute magnitude $M_v<0$ and an apparent magnitude limit of $m_v<23$.
- 10% error in distance and 10km/s of error in velocity was added to this data. For [Fe/H] and [α/Fe] an error of 0.1 dex was added.
- Starting with applying the group finder in the 3d space of x,y,z co-ordinates we progressively add extra dimensions in the form of radial velocity v_r , metallicity [Fe/H] and alpha element abundance [α/Fe] and finally the three velocity components v_x v_y v_z .
- A plot of n_g the number of stars in the group and precision P_r shown on the right demonstrates that both the number of groups recovered and their precision increases as extra dimensions are added.



GAIA Survey

- We generated a synthetic survey of giants stars $M_v<0$ and having an apparent magnitude limit of $m_v<18$. The distance, radial velocity and proper motion errors were added in accordance with the expected errors for the GAIA mission. The results of applying the group finder on 11 such surveys is shown on the right. In 6d position and velocity space about 37groups are recovered and these belong to 8 unique satellites on average.



Discussion and Conclusions

- For group finding in higher dimensions it is important to calculate the distance metric appropriately.
  - The metric should be locally adpative and anisotropic in general and uninformative dimensions needs to be scaled properly.
  - The algorithm proposed here utilising Shannon entropy and eigen decomposition can be used to calculate such a metric.
- Signatures of hierarchical structure formation should manifest itself in the stellar halo as substructures in either the phase space or abundance space. Using multidimensional group finding techniques these can be identified.
  - The more the amount of information in the form of dimensions, the easier it is to identify groups.