LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Faculty of Physics • Philipp Frank

SOMBI - Bayesian identification of parameter relations in cosmological data

Interdisciplinary Cluster Workshop:
"Challenges in statistical inference"

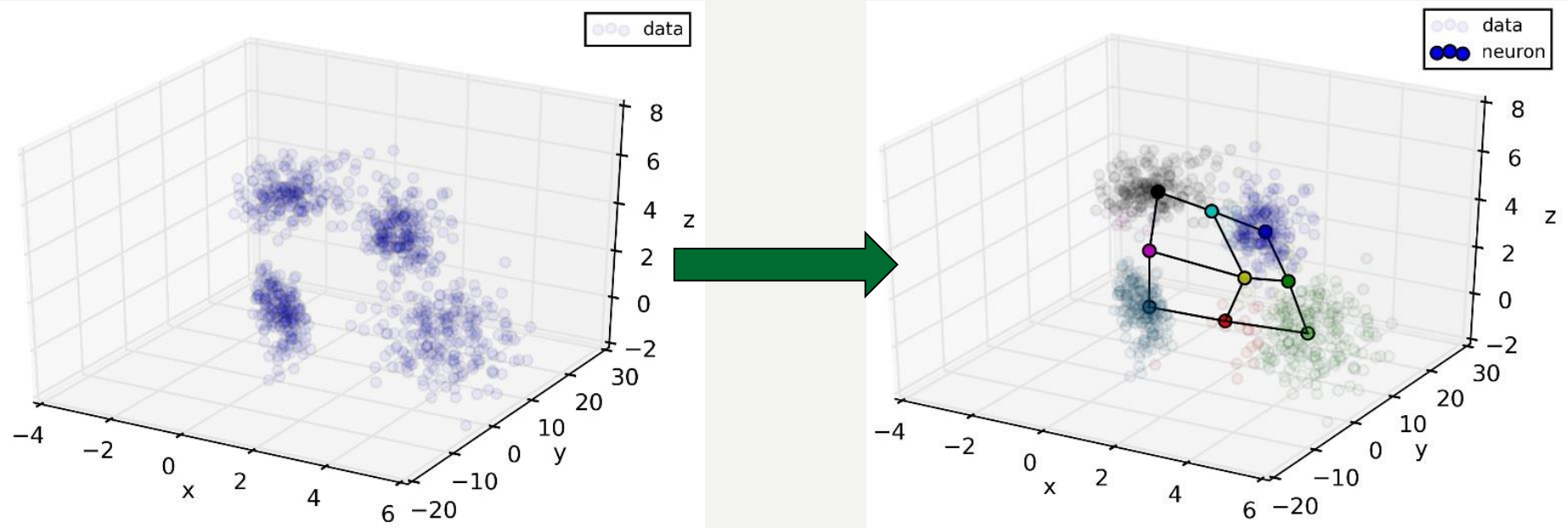
Motivation

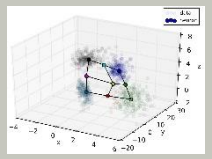
- Correlation between two quantities (x, y)
- Part of a large, high-dimensional and highly structured dataset
- Combination of data drawn from multiple generation processes

➢ Sub-samples of data holding various different correlation structures

  ➢ Assumed to be spatially separated in the dataspace

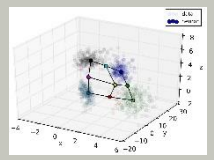
Self organizing map (SOM)





SOM - Setup

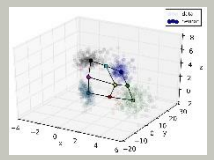
- Neuron space: regular grid of points (neurons)
- Each neuron holds a data space vector: weight $\mathbf{W}$
- Recursive update rule for all $\mathbf{W}$
- Conditional to the data $D = \{\mathbf{V}_t \ , \ t \in (1, .., N)\}$
  - ➢ Iterative learning process



SOM - Iteration step

- Calculate the best matching unit (BMU) for $\mathbf{V}_t$, which is the closest neuron to $\mathbf{V}_t$
- Normalized Euclidean metric: $Dist = \sqrt{\sum_{i=1}^{N} \left(\frac{V_t^i - W^i}{\sigma_i} \right)^2}$

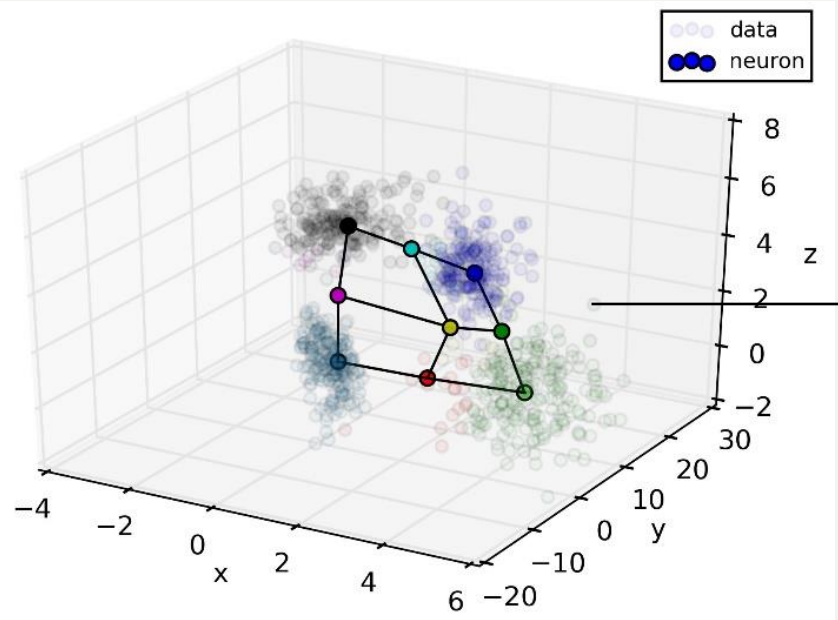
  $\sigma_i := V_{\max}^i - V_{\min}^i$
- New weight for BMU: $\mathbf{W}_{t+1} = \mathbf{W}_t + L_t \left(\mathbf{V}_t - \mathbf{W}_t \right)$
- Learning rate: $L_t = L_0 \, e^{-\frac{t}{\lambda}} \quad L_0 \in [0,1], \lambda \in \mathbb{R}^+$

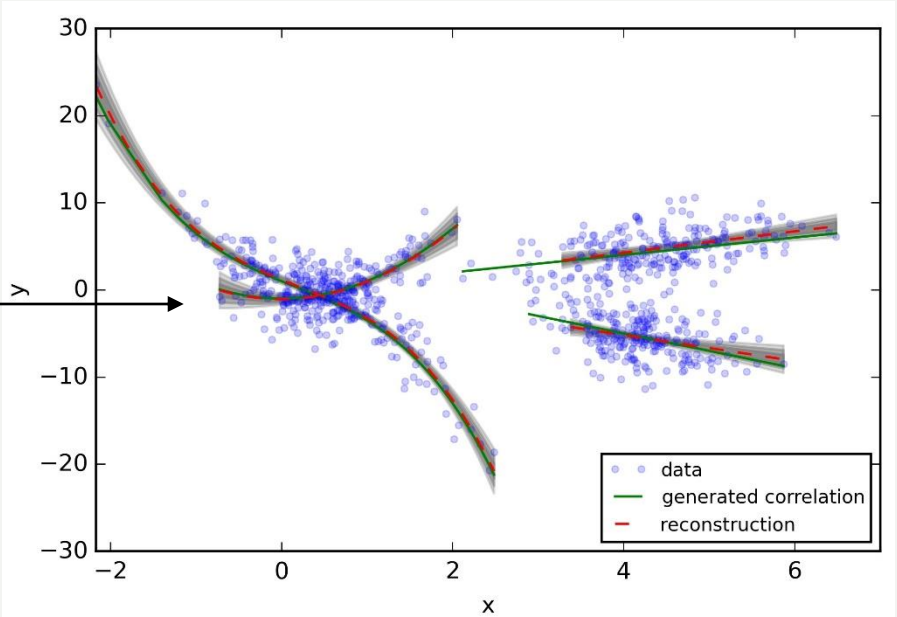


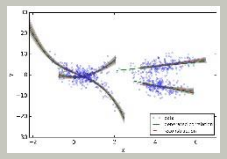
SOM - Iteration step

- Update other weights $\mathbf{W}_{t+1} = \mathbf{W}_t + L_t \Theta_t \, (\mathbf{V}_t - \mathbf{W}_t)$
- Neighbourhood function $\Theta_t = e^{-\frac{(d_{\mathrm{BMU}})^2}{2\sigma_t^2}}$
- Neighbourhood size $\sigma_t = \sigma_0 \, e^{-\frac{t}{\lambda}} \quad \sigma_0, \lambda \in \mathbb{R}^+$

Correlation analysis







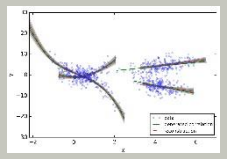
Correlation analysis

- Sub-sample of data: $\mathbf{d} = \{(x_j, y_j), \quad j \in (1, ..., U)\}$
- Model: $y_j = f(x_j) + n_j \approx \sum_{i=0}^{M} f_i(x_j)^i + n_j$

$$\mathbf{y} = \mathbf{Rf} + \mathbf{n} =$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_U \end{pmatrix} = \begin{pmatrix} 1 & x_1^1 & x_1^2 & \dots & x_1^M \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_U^1 & x_U^2 & \dots & x_U^M \end{pmatrix} \begin{pmatrix} f_0 \\ f_1 \\ \dots \\ f_M \end{pmatrix} + \begin{pmatrix} n_1 \\ n_2 \\ \dots \\ n_U \end{pmatrix}$$

SOMBI - Correlation analysis



- Noise: $P(\mathbf{n}|\eta) := \mathcal{G}(\mathbf{n}, \mathbf{N}) \quad \mathbf{N} = e^{\eta}\mathbb{1}$
- Priors: $P(\eta) = const. \qquad P(\mathbf{f}) = const.$

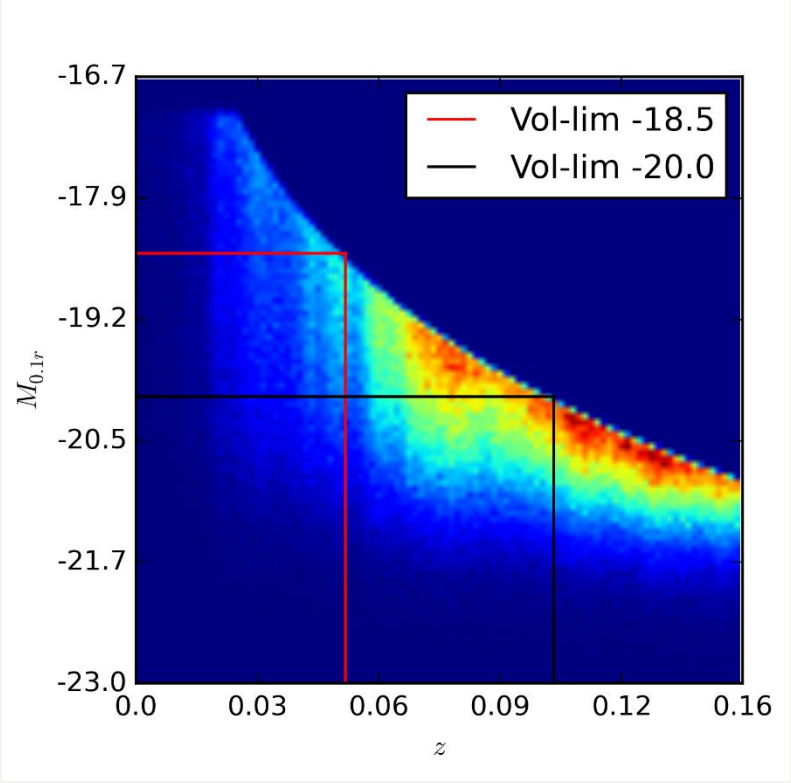
➢ Posterior:
$$P(\mathbf{f}|\mathbf{d}, \eta) = \frac{P(\mathbf{f}, \mathbf{d}, \eta)}{\int P(\mathbf{f}, \mathbf{d}, \eta)\ \mathrm{d}\mathbf{f}} = \mathcal{G}(\mathbf{f} - \mathbf{D}\mathbf{j}, \mathbf{D})$$
$$\mathbf{D} = (\mathbf{R}^T\mathbf{N}^{-1}\mathbf{R})^{-1} \quad \mathbf{j} = \mathbf{R}^T\mathbf{N}^{-1}\mathbf{y}$$

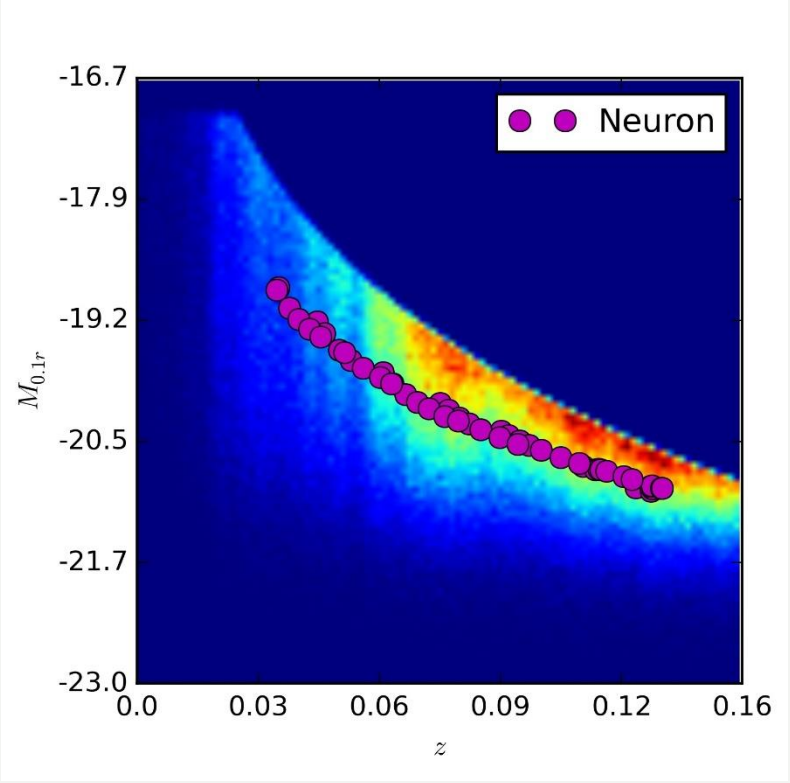
- Estimate for η: Empirical Bayes
- $$P(\eta|\mathbf{d}) \propto \int P(\mathbf{f}, \mathbf{d}, \eta)\ \mathrm{d}\mathbf{f} = \frac{|2\pi\mathbf{D}|^{\frac{1}{2}}}{|2\pi\mathbf{N}|^{\frac{1}{2}}}\ e^{-\frac{1}{2}(\mathbf{y}^T\mathbf{N}^{-1}\mathbf{y} - \mathbf{j}^T\mathbf{D}\mathbf{j})}$$
- Maximum a posterior: $$e^{\eta_{\mathrm{MAP}}} = \frac{\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{R}(\mathbf{R}^T\mathbf{R})^{-1}\mathbf{R}^T\mathbf{y}}{U - (M+1)}$$

Cosmological application

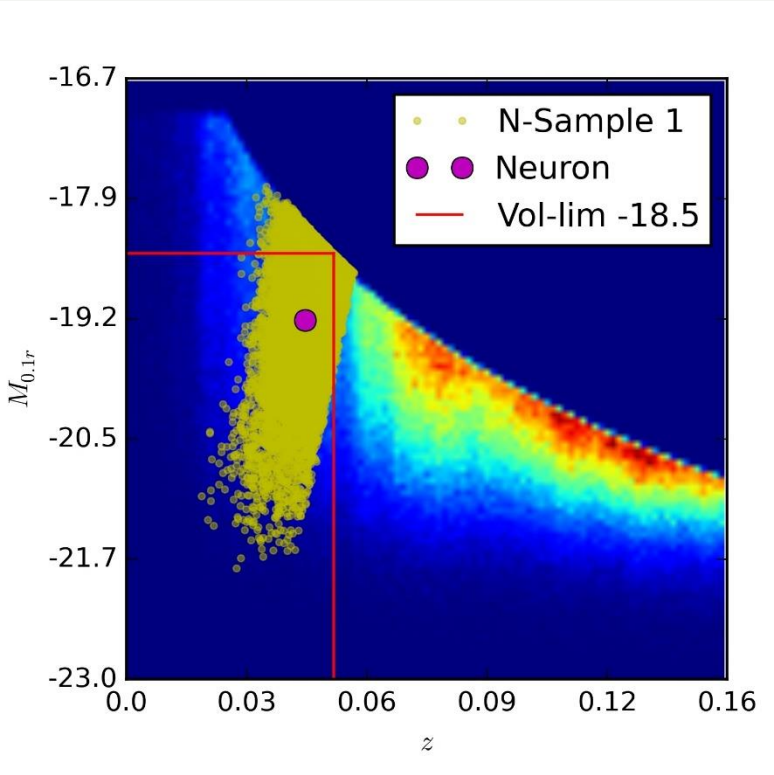
- Correlation analysis between:
  - Galaxy data (stellar mass, absolute magnitude, color, …)
    - SDSS DR7 main sample (Abazajian et al. 2009)
  - Cosmic Large-scale-structure (LSS) (density field, tidal-shear tensor of gravitational Potential, …)
    - Borg algorithm (Jasche & Wandelt 2013)

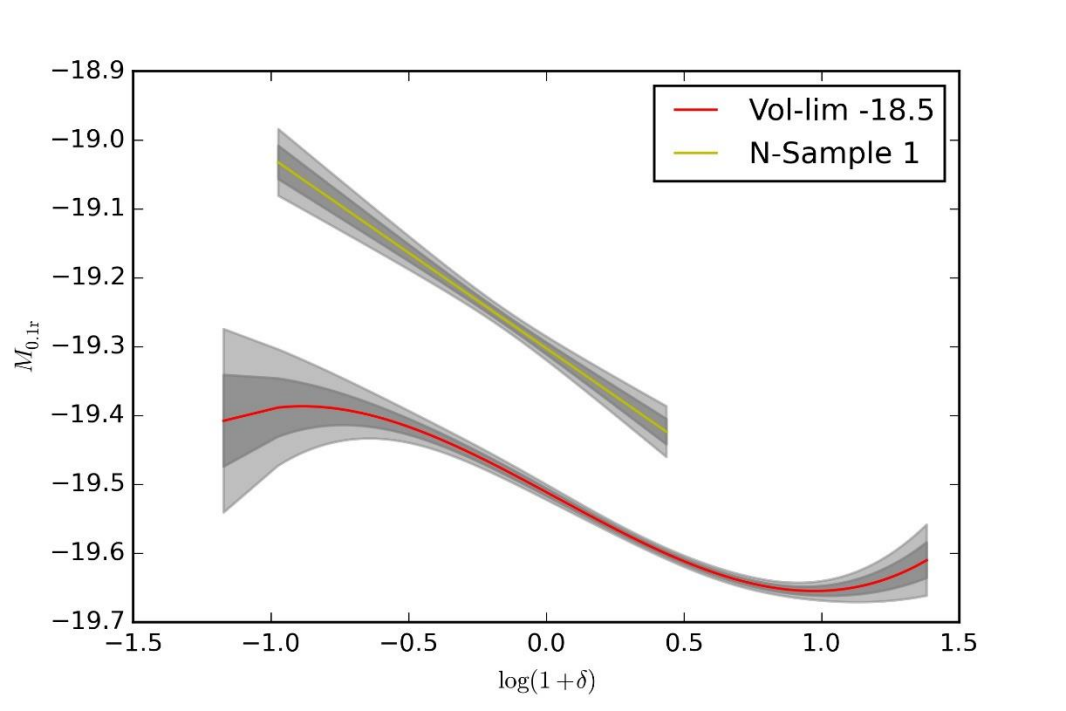
Known systematics



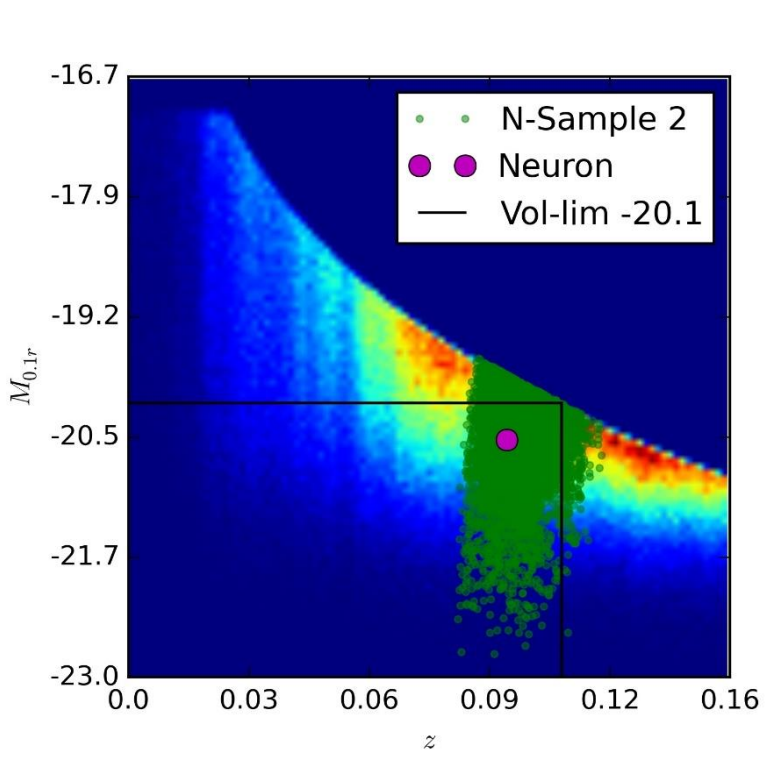


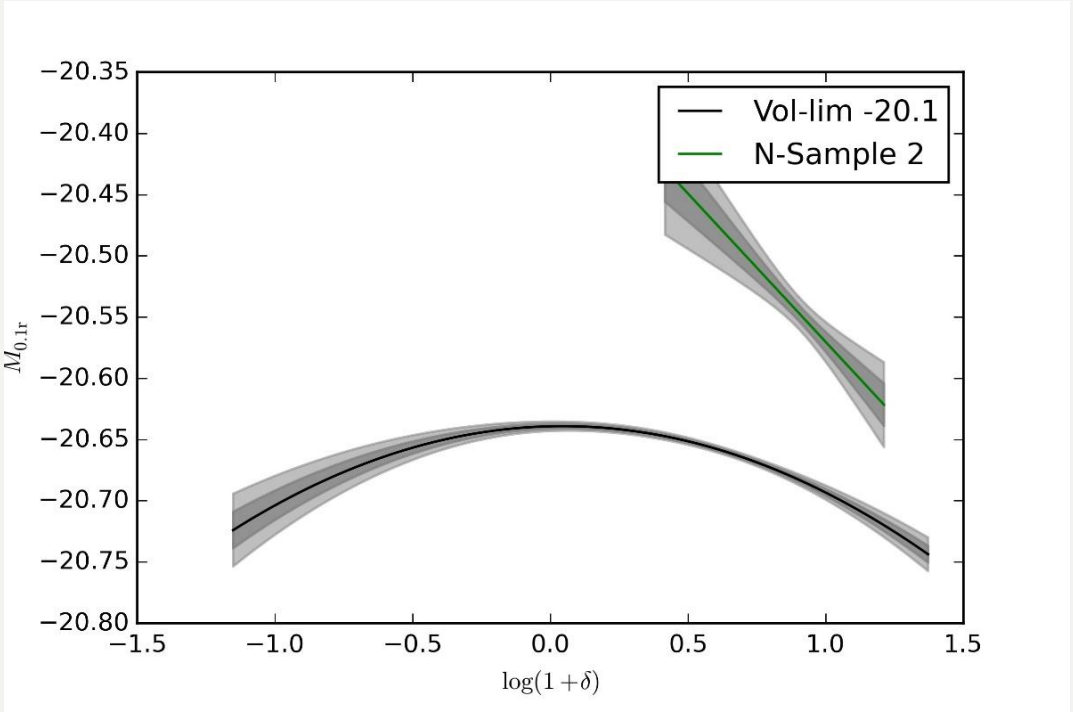
SOMBI - Comparison



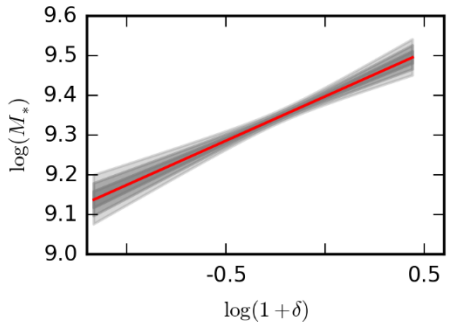


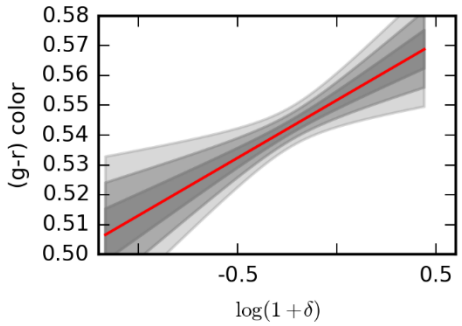
SOMBI - Comparison

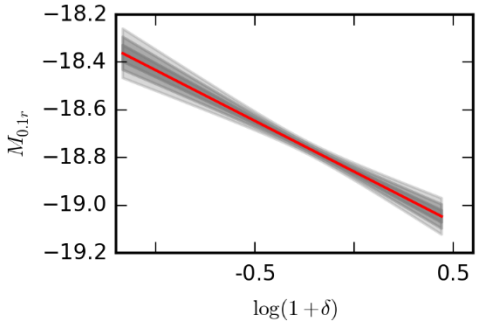




SDSS main sample



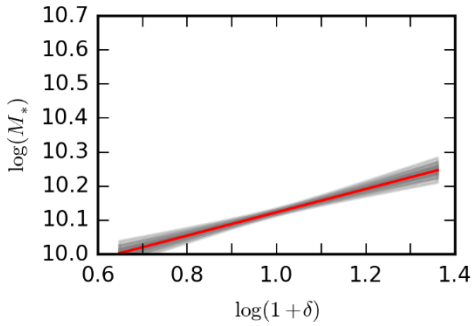


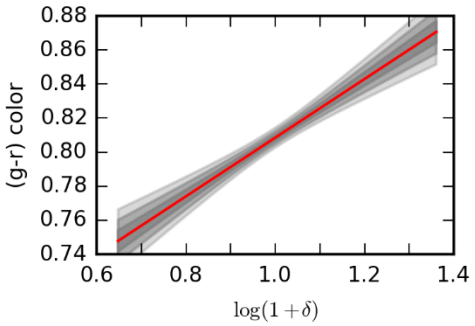


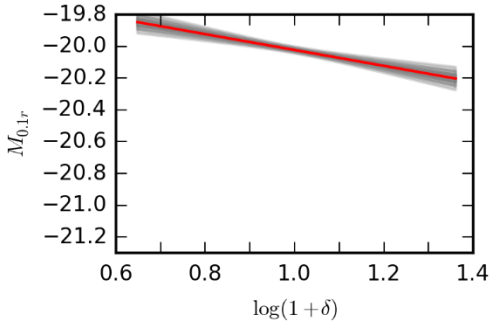
Data space weight W of BMU

z	$(g-r)color$	$M_{0.1r}$
0.04	0.59	-19.07
$\log(1+\delta)$	λ_1	λ_2
-0.06	-0.57	-0.0
λ_3		
0.63		

SDSS main sample



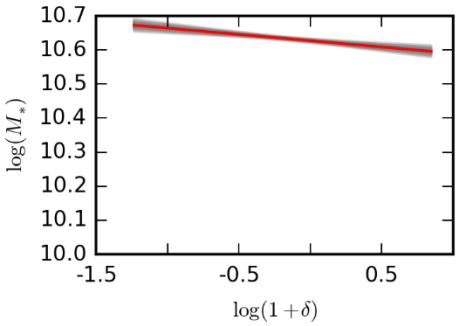


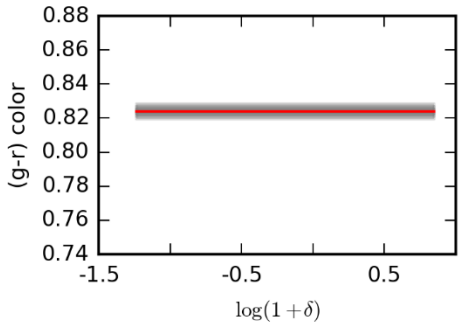


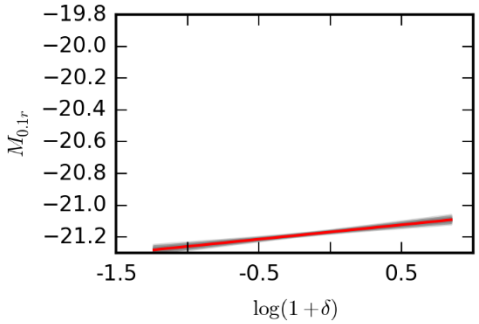
Data space weight W of BMU

z	$(g-r)color$	$M_{0.1r}$
0.07	0.79	-19.96
$\log(1+\delta)$	λ_1	λ_2
0.9	0.62	2.61
λ_3		
4.31		

SDSS main sample







Data space weight W of BMU

z	$(g-r)color$	$M_{0.1r}$
0.14	0.81	-21.04
$\log(1+\delta)$	λ_1	λ_2
-0.02	-0.58	0.09
λ_3		
0.79		