Carnegie Mellon University
University of Pennsylvania
January 2012

Cluster Analysis, Model Selection, and Prior Distributions on Models

George Casella
University of Florida

Elías Moreno
University of Granada

F. Javier Girón
University of Málaga

Introduction

The Clustering Problem

- $Y \sim \mathfrak{F} = \{f(y|\theta), \theta \in \Theta\}$, where $\Theta \in \mathbb{R}^k$
- We observe a sample of n independent data $\mathbf{y} = (y_1, y_2, ..., y_n)$
- We look at the sample as being split into clusters,
  - ▷ Observations within a cluster are from the same sample density $f(y|\theta)$
  - ▷ The parameter θ of the density changes across clusters.
- Goal: To reduce the number of sampling models M_j
  - ▷ By clustering the observation coming from the same model
- This is, in fact, a Model Selection problem.

Outline of the Talk

- Some Background — Challenges, Models
- Structure of Clustering — Classifying the clusters
- Priors on Models — Uniform? or something else?
- Consistency — Not all priors are equal
- Bayes Factors — Intrinsic Priors
- Implementation — Searching and Clustering Regressions
- Conclusions — How to cluster

Background

Just How Many Clusters are there in the Galaxy Data?

- Galaxy Data from Postman *et al.* (1986): measurements of velocities in 10^3 km/sec of 82 galaxies from a survey of the Corona Borealis region.
- Roeder (1990): at least 3, no more than 7 modes (Confidence set)
- Others are in consensus

9172	9350	9483	9558	9775	10227
10406	16084	16170	18419	18552	18600
18927	19052	19070	19330	19343	19349
19440	19473	19529	19541	19547	19663
19846	19856	19863	19914	19918	19973
19989	20166	20175	20179	20196	20215
20221	20415	20629	20795	20821	20846
20875	20986	21137	21492	21701	21814
21921	21960	22185	22209	22242	22249
22314	22374	22495	22746	22747	22888
22914	23206	23241	23263	23484	23538
23542	23666	23706	23711	24129	24285
24289	24366	24717	24990	25633	26960
26995	32065	32789	34279		

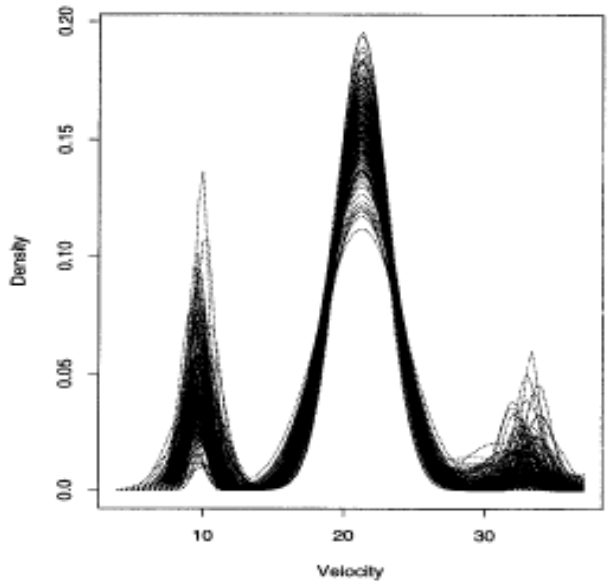


Figure 1. Densities Obtained From the Markov Chain Monte Carlo Sampler Using the Astronomy Data From Roeder (1992).

- Histogram from Roeder and Wasserman (1997)

Background

Modes/Clusters in the Galaxy Data: The Statistics All-Star Team

Roeder (1990)	at least 3, no more than 7 (Confidence set)
Richardson & Green (1997)	6 has highest posterior probability
Roeder & Wasserman (1997)	The posterior clearly supports three groups
Lau & Green (2007)	Optimal number of clusters is three
Wang & Dunson (2011)	Five clusters

- Anyone want to bet that there are more than SEVEN??

Background

Clusters in the Galaxy Data: A Simple Approach

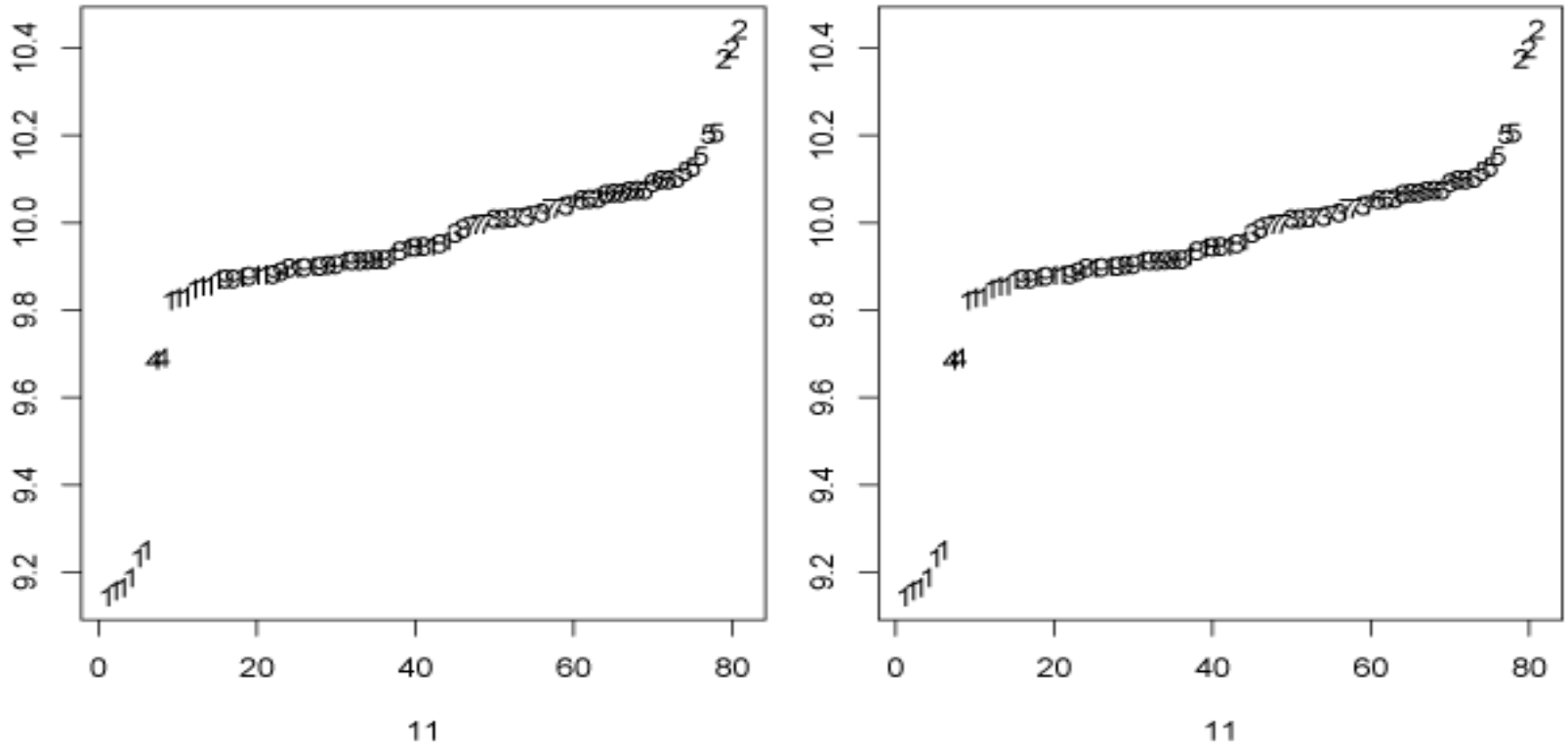
- Measure partition using BIC (or any other Model Selector)
- Prior distribution says all models equally likely
- Run a stochastic search to maximize BIC
  - Compare current partition to one-cluster
  - $50,000$ iterations (more if you like)
- Reasonable?

Background

Clusters in the Galaxy Data: What BIC Found

- Top partitions had 11 clusters
- What Happened?

Rank	# Clusters	BIC $\times 10^{71}$
1	11	49.53
2	11	33.13
3	11	20.19
4	11	15.48
5	11	14.39



Background
Challenges

- Assuming that a family of sampling models $\mathfrak{F}$ has been chosen
- Need to assess the prior distribution for the discrete parameters
- Need to assess the prior for the (usually) continuous parameters θ
  - The densities inside the clusters.
- We typically lack substantive prior information on these parameters
  - This leads us to propose the use of objective priors.

- Also, we need to compute the very many posterior model probabilities
  - There are very many of them

Background

Product Partition Models

- Product Partition Models (PPM):

  ▷Hartigan (1990) ▷ Barry and Hartigan (1992)
  ▷Quintana and Iglesias (2003) ▷ Booth *et al.* (2008)

- Product Partition Formulation

  ▷ The sampling density of the data $\mathbf{y}$ conditional on a partition $\mathbf{r}_p$ is

$$f(\mathbf{y}|p, \mathbf{r}_p, \theta_p) = \prod_{i=1}^{n} f(y_i|\theta_{r_i^{(p)}}).$$

1. New observations classified to maximize probability
2. We can compute the posterior probability of any given partition.

Structure of Clustering
Introduction

- $\mathbf{r}_p = (r_1^{(p)}, ..., r_n^{(p)})$ is a *partition* of the sample into p clusters
  - $r_i^{(p)}$, $i = 1, \ldots n$ is an integer between 1 and p
  - y_i is assigned to cluster $r_i^{(p)}$

- For $n = 4$
  - There are 15 possible partitions (models), the *Bell number* for $n = 4$.
  - In each of the cluster classes, $p = 1, 2, 3, 4$ there are $1, 7, 6, 1$ partitions
  - $1, 7, 6, 1$ are *Stirling numbers of the second kind*

Structure of Clustering

Partitions for $n = 4$

$p = 1$	$p = 2$		$p = 3$	$p = 4$
$\mathfrak{R}_1$	$\mathfrak{R}_2$		$\mathfrak{R}_3$	$\mathfrak{R}_4$
$y_1 y_2 y_3 y_4$	$y_1 \mid y_2 y_3 y_4$	$y_1 y_2 \mid y_3 y_4$	$y_1 \mid y_2 \mid y_3 y_4$	$y_1 \mid y_2 \mid y_3 \mid y_4$
	$y_2 \mid y_1 y_3 y_4$	$y_1 y_3 \mid y_2 y_4$	$y_1 \mid y_3 \mid y_2 y_4$	
	$y_3 \mid y_1 y_2 y_4$	$y_1 y_4 \mid y_2 y_3$	$y_1 \mid y_4 \mid y_2 y_3$	
	$y_4 \mid y_1 y_2 y_3$		$y_2 \mid y_3 \mid y_1 y_4$	
			$y_2 \mid y_4 \mid y_1 y_3$	
			$y_3 \mid y_4 \mid y_1 y_2$	

- Four Cluster Classes, Five Configuration Classes
- The number of configuration classes in each cluster class is $b(n, p)$
  - $b(n, p) =$ partitions of the integer n into p components ≥ 1
  - $b(4, 1) = 1, \quad b(4, 2) = 2, \quad b(4, 3) = 1, \quad b(4, 4) = 1,$
- *Four Cajas, Five Cajitas*

Structure of Clustering

Models

- The sampling density of the data $\mathbf{y}$ conditional on a given partition $\mathbf{r}_p$ is

$$f(\mathbf{y}|p, \mathbf{r}_p, \theta_p) = \prod_{i=1}^{n} f(y_i|\theta_{r_i^{(p)}}).$$

- The class of models is $\mathfrak{M} = \cup_{1 \leq p \leq n} \mathfrak{M}_p$
  - Models in $\mathfrak{M}_p$ have p clusters

- The generic Bayesian model is given as

$$M_{\mathbf{r}_p} : \{f(\mathbf{y}|p, \mathbf{r}_p, \theta_p),\ \pi(p, \mathbf{r}_p, \theta_p|n)\}.$$

- We'll talk about the prior later...

Structure of Clustering
Bayes Factors

- We use as a reference the partition with one cluster, $M_{\mathbf{r}_1}$.
- We can write the posterior probability of $M_{\mathbf{r}_p}$ in the class $\mathfrak{R}_p$ as

$$\pi(\mathbf{r}_p|\mathbf{y},p,n) = \frac{\pi(p,\mathbf{r}_p|n)B_{\mathbf{r}_p,\mathbf{r}_1}(\mathbf{y})}{\sum_{\mathbf{r}_p\in\mathfrak{R}_p}\pi(p,\mathbf{r}_p|n)B_{\mathbf{r}_p,\mathbf{r}_1}(\mathbf{y})}, \quad \text{if} \quad \mathbf{r}_p \in \mathfrak{R}_p,$$

- $B_{\mathbf{r}_p,\mathbf{r}_1}(\mathbf{y})$ is the Bayes factor for comparing model $M_{\mathbf{r}_p}$ against $M_{\mathbf{r}_1}$.
- In the class of all models the posterior probability of model $M_{\mathbf{r}_p}$ is

$$\pi(\mathbf{r}_p|\mathbf{y}) = \frac{\pi(p,\mathbf{r}_p|n)\ B_{\mathbf{r}_p,\mathbf{r}_1}(\mathbf{y})}{\sum_{p=1}^{n}\sum_{\mathbf{r}_p\in\mathfrak{R}_p}\pi(p,\mathbf{r}_p|n)\ B_{\mathbf{r}_p,\mathbf{r}_1}(\mathbf{y})}, \ \text{if} \ \mathbf{r}_p \in \mathfrak{R}$$

Prior Distributions on Models
Introduction

- We factor the prior for models as

$$\pi(p, \mathbf{r}_p|n) = \pi(\mathbf{r}_p|p, n)\pi(p|n).$$

- $\pi(\mathbf{r}_p|p, n)$ is much more important than $\pi(p|n)$
  - Both factors depend on n
  - $\pi(\mathbf{r}_p|p, n)$ is much more sensitive to n
  - The size of the cluster classes grows exponentially with n

- Uniform Prior A popular choice, giving equal probability to all models:

$$\pi^U(p, \mathbf{r}_p|n) = \frac{1}{\mathcal{B}_n}, \quad \mathcal{B}_n \text{ is the Bell number.}$$

  - This seemingly innocuous choice can have unforeseen consequences.

Prior Distributions on Models
The Hierarchical Uniform Prior (HUP)

- An objective prior for models,using the structure of the cluster problem
- To carry our our prior specification, we start from the decomposition

$$\pi(p, \mathbf{r}_p|n) = \underbrace{\pi(\mathbf{r}_p|\mathfrak{R}_{p;n_1,\ldots,n_p}, n)}_{\text{Model}} \underbrace{\pi(\mathfrak{R}_{p;n_1,\ldots,n_p}|p, n)}_{\text{Configuration}} \pi(p|n)$$

- The partitions $\mathbf{r}_p$ in $\mathfrak{R}_{p;n_1,\ldots,n_p}$ assign n_i components to $f(\cdot|\theta_i)$.
  - ▷ We assign a uniform prior

$$\pi(\mathbf{r}_p|\mathfrak{R}_{p;n_1,\ldots,n_p}, n) = \text{Uniform in } \mathfrak{R}_{p;n_1,\ldots,n_p} \textit{ (Uniform inside Cajitas)}$$

- Assume that the partitions $\mathfrak{R}_{p;n_1,\ldots,n_p}$ in $\mathfrak{R}_p$ *a priori* equally likely, so

$$\pi(\mathfrak{R}_{p;n_1,\ldots,n_p}|p, n) = b(n, p)^{-1} \quad \textit{(Uniform among Cajitas)}$$

  - ▷ And the hierarchical uniform specification is complete

Prior Distributions on Models
Comparisons

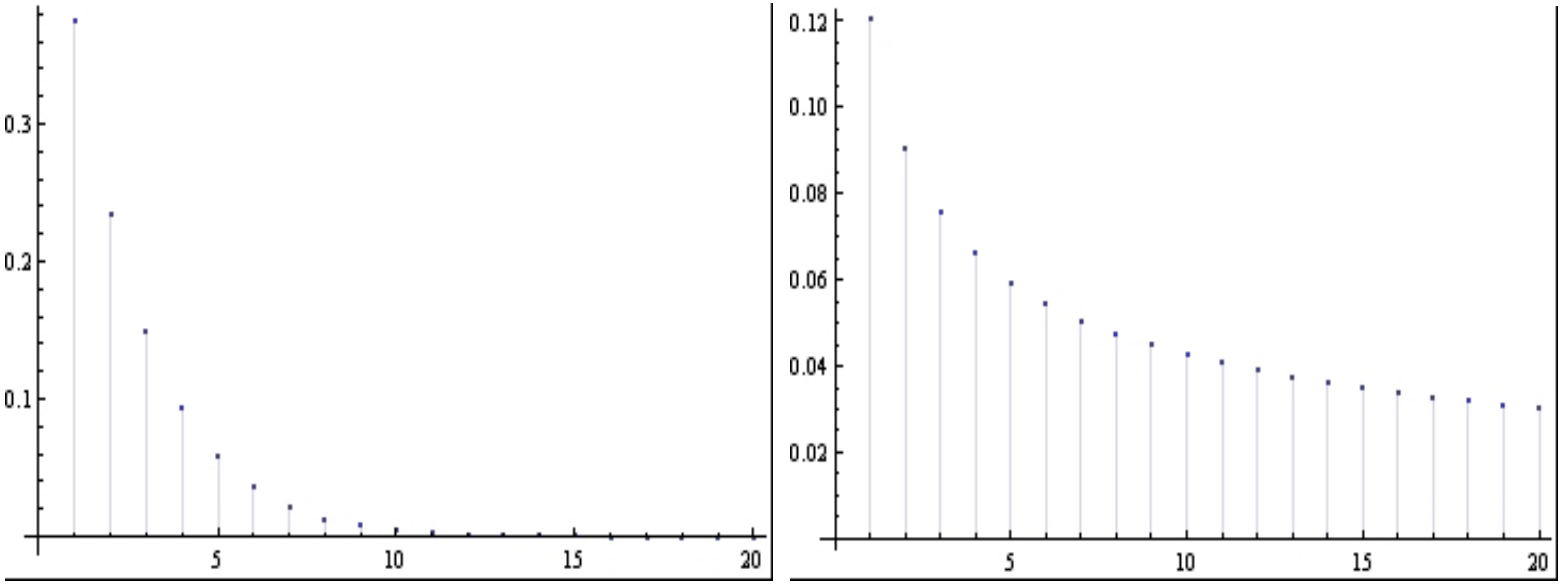
- A small numerical example
- Prior probabilities for exchangeable partition sets in $\mathfrak{R}_3$ for $n = 10$.

Configuration	Hierarchical Uniform	Uniform
$\{1, 3, 6\}$	0.14	0.09
$\{2, 3, 5\}$	0.14	0.27

- The partitions are very close, only differing by moving one observation.
  - ▷ Only the HUP gives these partitions the same prior probability
- *A priori* it does not seem that, for $n = 10$, we would have any reason to assign different probabilities to the configurations $\{1, 3, 6\}$ and $\{2, 3, 5\}$
- $\#\{1, 3, 6\} = 840$ $\#\{2, 3, 5\} = 2525$

Prior Distributions on Models
Prior Distributions for the Number of Clusters

- For the prior $\pi(p, \mathbf{r}_p|n)$ we need to specify $\pi(p|n)$.
- We desire a relatively small number of clusters in the sample
  - The extreme case of n clusters should be given a very small probability.



- Left Panel: Poisson-Intrinsic
- Right Panel: Poisson-Jeffreys

The Role of the Prior in Consistency of the Bayes Procedure

Introduction

- Bayesian model selection is consistent when
  - The dimension of the sampling model is fixed
  - Comparisons are pairwise
- Follows from consistency of the Bayes factor
  - The model prior does not play any role in the consistency
- However, when the dimension of the model grows with the sample size
  - The model prior plays an important role for obtaining consistency
  - This is the case in clustering.
- Surprisingly, the actual choice of Bayes factor
  - Is of almost no consequence in determining consistency
  - Many Bayes factors have the same asymptotic representation.

The Role of the Prior in Consistency of the Bayes Procedure

Consistency and Inconsistency

- As $n \to \infty$
  - Consistency is specific to the limiting configuration $\left(\frac{n_1}{n}, \ldots, , ..., \frac{n_p}{n}\right)$
  - Assume that the number of clusters p is bounded, that is, $p \leq M < \infty$.
- Theorem With the uniform prior, when sampling from $M_{\mathbf{r}_1}$,

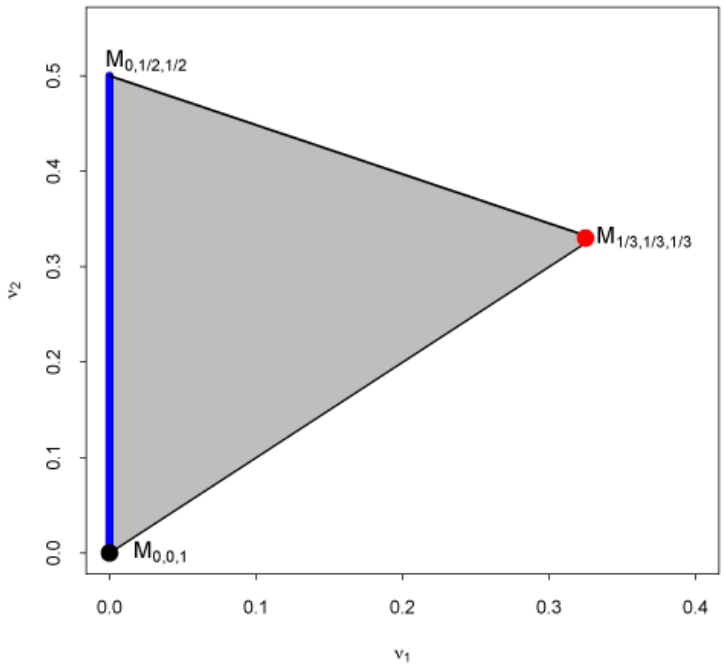
$$\lim_{n\to\infty} \Pr(\mathfrak{R}_p|\mathbf{y}) = \begin{cases} 1, & \text{if} \quad p = M, \\ 0, & \text{if} \quad p \leq M-1. \end{cases}$$

  - The uniform prior picks the biggest model with probability one.
  - It is also the wrong model.
- Theorem With the hierarchical uniform prior, when sampling from $M_{\mathbf{r}_1}$,

$$\lim_{n\to\infty} \Pr(\mathfrak{R}_1|\mathbf{y}) = 1.$$

  - The correct model is chosen with probability 1.

The Role of the Prior in Consistency of the Bayes Procedure
Limiting Distributions ($p = 3$)



- $\mathfrak{R}_1 =$ Black Dot
- $\mathfrak{R}_2 =$ Blue Line
- $\mathfrak{R}_3 =$ Grey Triangle

- The HUP converges in distribution to a uniform prior on the simplex.
- The Uniform Prior converges distribution to a degenerate distribution
  - Concentrated at the vertex $(\frac{1}{p}, \ldots, \frac{1}{p})$ of the simplex
  - Red Dot

Intrinsic Priors for the Continuous Parameters
Linear Models

- Suppose that the sample $(y_1, ..., y_n)$ follows a normal linear model

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon, \quad \varepsilon \sim N_n(\varepsilon|\mathbf{0}, \tau^2\mathbf{I}_n),$$

- Since $f(\mathbf{y}|1, \mathbf{r}_1, \beta, \tau)$ is nested in $f(\mathbf{y}|p, \mathbf{r}_p, \beta_1, .., \beta_p, \sigma_p)$,
  - ▷ Intrinsic methodology gives the intrinsic prior for the parameter $(\beta_1, ..., \beta_p, \sigma_p)$
- The Intrinsic Bayes factor is a function of

$$\mathcal{R}_{\mathbf{r}_p} = \frac{RSS_{n_1} + \ldots + RSS_{n_p}}{RSS_n}, \quad RSS_{n_i} = \text{ residual sum of squares}$$

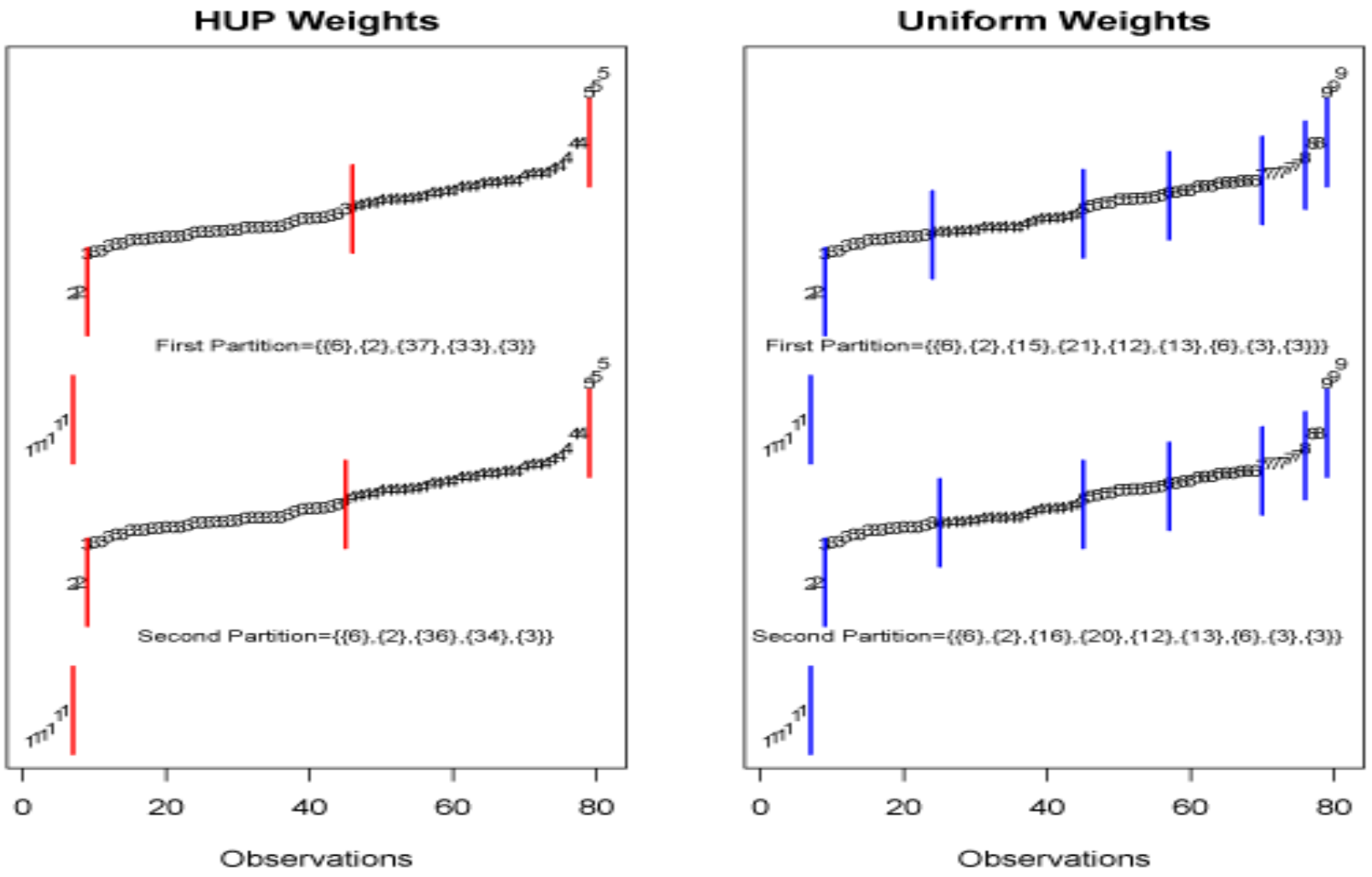
  - ▷ Along with an ugly dimension correction

- Hybrid search algorithm, based on a Metropolis-Hastings algorithm
- Variation of the biased random walk of Booth *et al.* (2008).

Implementation
Galaxy Data

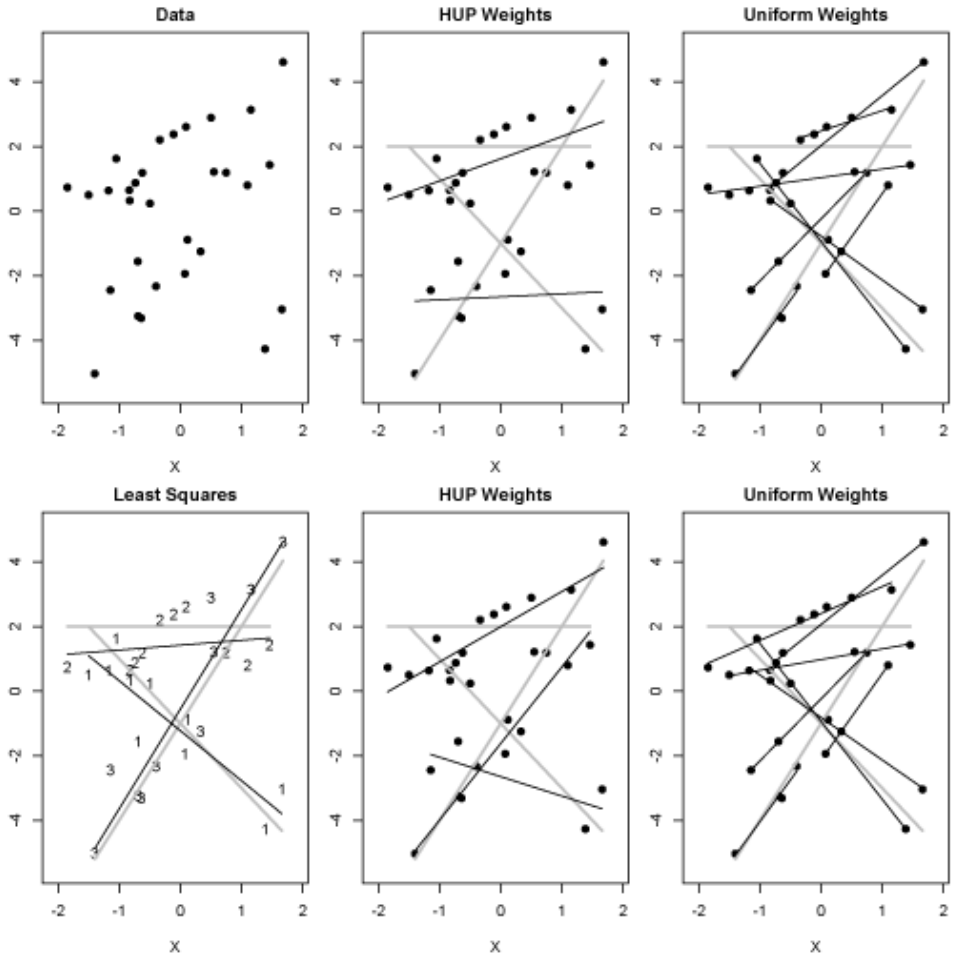
- We return Galaxy data (Roeder 1990) as a benchmark.
- Observations on the velocity (km/second) of 81 galaxies.
  - Recall: It is well accepted that there are between 5 and 7 clusters in the data
- The top 25 HUP partitions in the search all had 5 clusters.
- The Uniform search found partitions with 9 clusters
  - By consensus this is too many clusters.

Galaxy Data



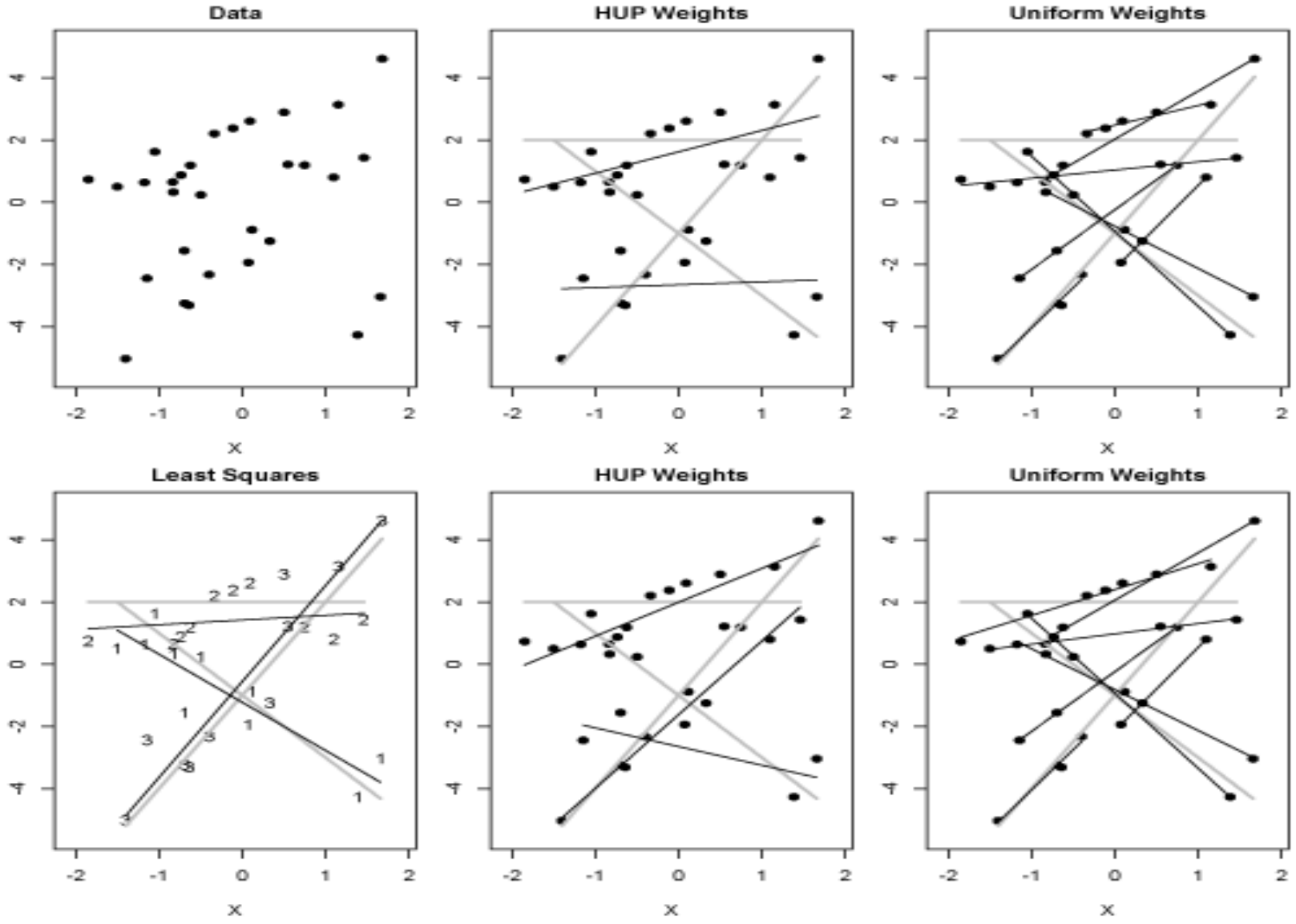
- HUP top and bottom clusters differ only by one shift point in the middle string
  - Y-axis is log(Velocity)

Evaluating the Procedures
Simulated Regression Data



- Left top: Simulated Data from three models
- Left bottom: True models (grey) and the least squares fit.
- Middle: Typical results from HUP weights
- Right: Typical results from uniform weights.
- A Bigger Picture...

Simulated Regression Data



Evaluating the Procedures
Concha y Toro Data

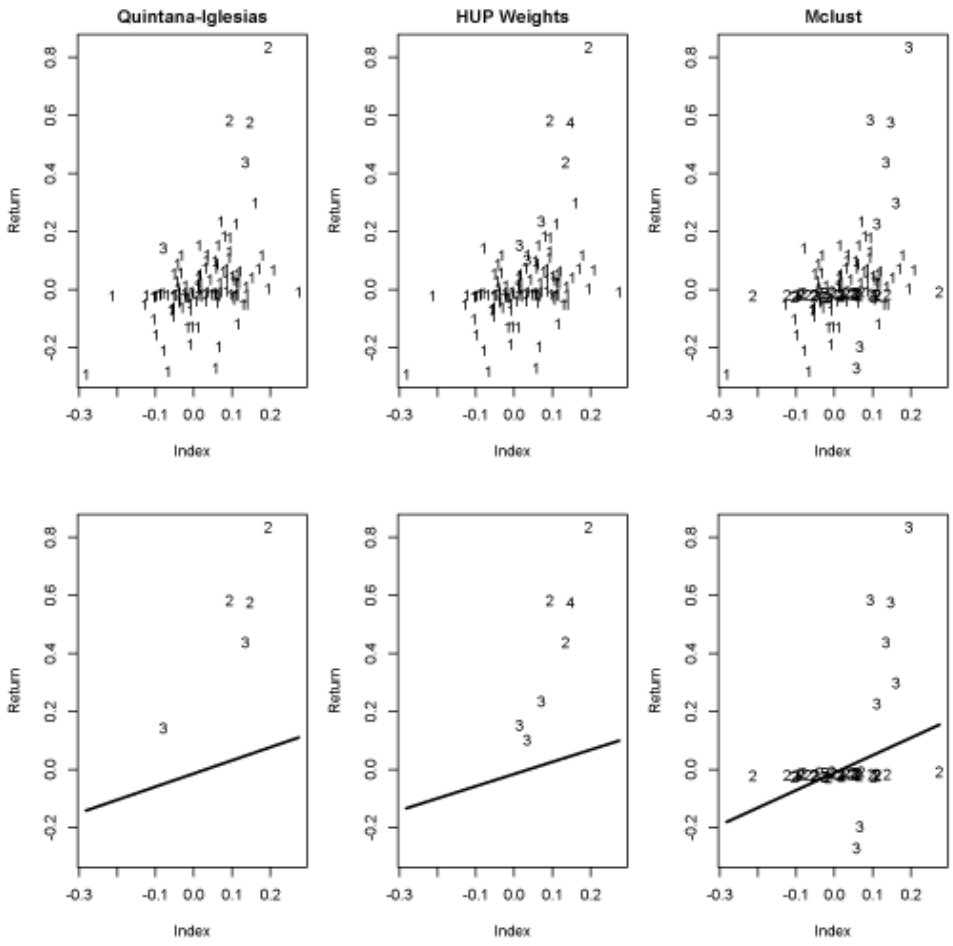
- Quintana and Iglesias (2003) (QI) analyze economic data pertaining to the winemaker *Concha Y Toro*.
- This is simple linear regression data, using a model of the form

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

  - y = the *Concha Y Toro* stock return
  - x = a stock market index, similar to the US Dow-Jones Index.

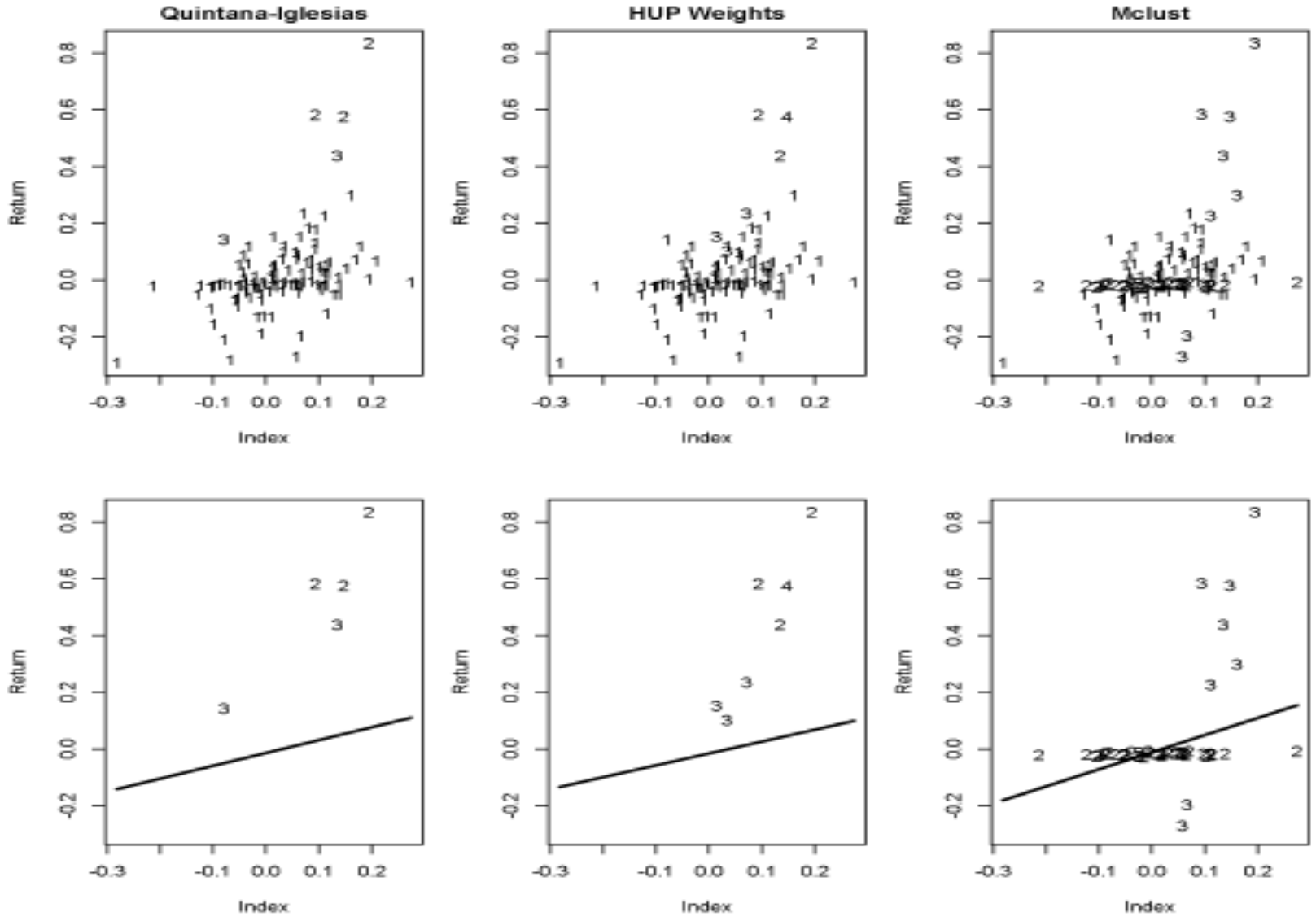
- QI use a version of their full PPM model set up for outlier detection
- We ran the data using only default settings and HUP weights

Evaluating the Procedures
Concha y Toro Data



- We found three clusters
  - ▷ one large one containing the data without outliers
  - ▷ Two other "outlier clusters"
- Similar to QI findings
- `Mclust` also found three clusters.
  - ▷ Two large, one small
- A Bigger Picture...

Concha y Toro Data



Conclusions

How to Cluster?

- ▶ Priors
  - HUP: Good and consistent
  - Uniform: Too many clusters and inconsistent
- ▶ Inference
  - Want small number of clusters
  - Need the prior to pull there
  - Even if the truth is a large number of clusters
- ▶ Limit Results
  - Emphasizes the value of HUP
- ▶ Theory and Examples
  - Findings are compatible

Conclusions
Next?

- Generalizations

  - Limit results apply to Bayes factors
  - BIC is asymptotically equivalent intrinsic Bayes
  - Can extend to a wide class of priors.

- Next?

  - Other models?
  - Linear mixed
  - Discrete

Thank You for Your Attention

George Casella
casella@ufl.edu

Elías Moreno
emoreno@ugr.es

F. Javier Girón
fj_giron@uma.es

Selected References
All on my web page

- Booth, J.G., Casella, G. and Hobert, J.P. (2008). Clustering using objective functions and stochastic search. *J. R. Statist. Soc. B,* **70**, 1, 119–139.
- Casella, G. and Moreno, E. (2006) Objective Bayes Variable Selection. *Journal of the American Statistical Association* **101** 157-167.
- Casella, G., Girón, F.J., Martínez, M.L. and Moreno E. (2009). Consistency of Bayesian procedure for variable selection. *Annals of Statistics*, **37,** 3, 1207-1228.
- Girón, F.J., Moreno, E., Casella, G. and Martínez, M.L. (2010). Consistency of objective Bayes factors for nonnested linear models and increasing model dimension. *RACSAM* **104** (1), 61–71.
- Moreno, E., Bertolino, F. and Racugno,W. (1998). An intrinsic limiting procedure for model selection and hypothesis testing. *J. Amer. Statist. Assoc.*, **93**, 1451-1460.
- Moreno, E., Girón, F.J. and Casella, G. (2010). Consistency of objective Bayesian tests as the model dimension increases. *Annals of Statistics* **38** 1937-1952