Large scale clustering of simulated gravitational wave triggers using S-means and Constrained Validation methods.

L.R. Tang, H. Lei, S. Mukherjee, S. D. Mohanty
University of Texas at Brownsville

GWDAW12, Boston, December 13-16, 2007
LIGO DCC # LIGO-G070840-00-0

Introduction

LIGO data is seen to pick up a variety of glitches in all channels. Glitch analysis is a critical task in LIGO detector characterization. Several analysis methods (kleine Welle, Q-scan, BN event display, Multidimensional classification, NoiseFloorMon etc. have been in use to explain the source of the glitches. Trigger clustering is an important step towards identification of distinct sources of triggers and unknown pattern discovery.

The study presents two new techniques for large scale analysis of gravitational wave burst triggers. Traditional approaches to clustering treats the problem as an optimization problem in an open search space of clustering models. However, this can lead to over-fitting problems or even worse, non-convergence of the algorithm. The new algorithms address these problems. S-means looks at similarity statistics of burst triggers and builds up clusters that have the advantage of avoiding local minima. Constrained Validation clustering tackles the problem by constraining the search in the space of clustering models that are “non-splittable” –models in which centroids of the left and right child of a cluster (after splitting) are nearest to each other. These methods are demonstrated by using

This work is supported by NSF CREST grant to UTB 2007 and NSF PHY-055842 (2006).

CLUSTERING INTRODUCTION

- Clustering: unsupervised classification of data points. It is a well researched problem for decades with applications in many disciplines.
- New Challenges: Massive amount of data, many traditional algorithms not suitable.
- K-means: the most efficient clustering algorithm due to low complexity *O(nkl)*, high scalability and simple implementation

Weakness: i) sensitive to initial partition, solution: repeat K-means; ii) converge to local minima, local minima is acceptable in practice; iii) centroid sensitive to outliers solution: fuzzy K-means; iv) the number of cluster, K, *must specified in advance. Solution:* ***S-means****!*

S-MEANS: SIMILARITY DRIVEN CLUSTERING

Step 1. *Randomly* Initialize *K centroids (user can specify any starting K,   K=1 by default).*

Step 2. *Calculate the similarities from every point to every* centroid. Then, for any point, if the highest similarity to centroid *ci is >T, group it* to cluster *i, otherwise, add it to a new cluster (the K+ 1th cluster) and let K increase by 1*.

Step 3. *Update each* centroid with the mean of all member points. If one group becomes empty, remove its centroid and reduce *K by 1*. *Repeat the Step 2 and 3 until no new* cluster is formed and none of the centroids moves.

S-MEANS VS K-MEANS

- Similar to *K-means*, *but with big differences*.
- Major difference: in the second step, which basically groups all the points to a *new cluster* whose highest similarity to *existing centroids is below the given threshold*, *while K-means forces points into a group*, *even it shouldn’t be*. ***This makes big difference in output clusters***!
- *Minor difference: K increases by 1 as new clusters forms in each iteration, k decreases by 1 as some clusters become empty.*

EXPERIMENTS

Convergence

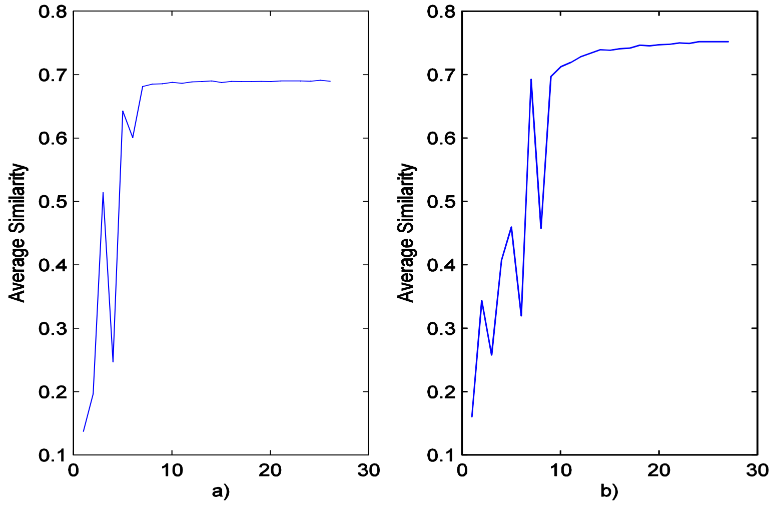


Fig. S-means converges on dataset SCT in less than 30 iterations. a) the maximum number of clusters is restricted (up-bound is set 6). b) without restriction (up-bound is set 600). Dataset used: UCI time series datasets, 600 x 60, 6 classes. Similarity used: correlation.

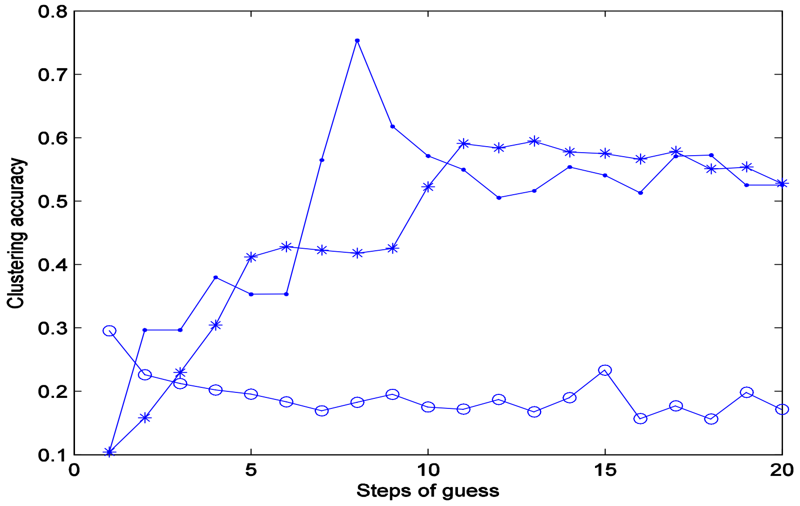


Fig. Accuracy comparison between S-means (dot point), fast *K-means (star point) and G-means*(circle point). *Confidence varies from 0.0001 to 0.002* in G-means. *T varies from 0 to 0.95 with step 0.05 and up-bound of clusters is set 20 in S-means.* Dataset used: PenDigit, 10992 x 32, 10 classes .

MINING GRAVITATIONAL-WAVE TRIGGERS

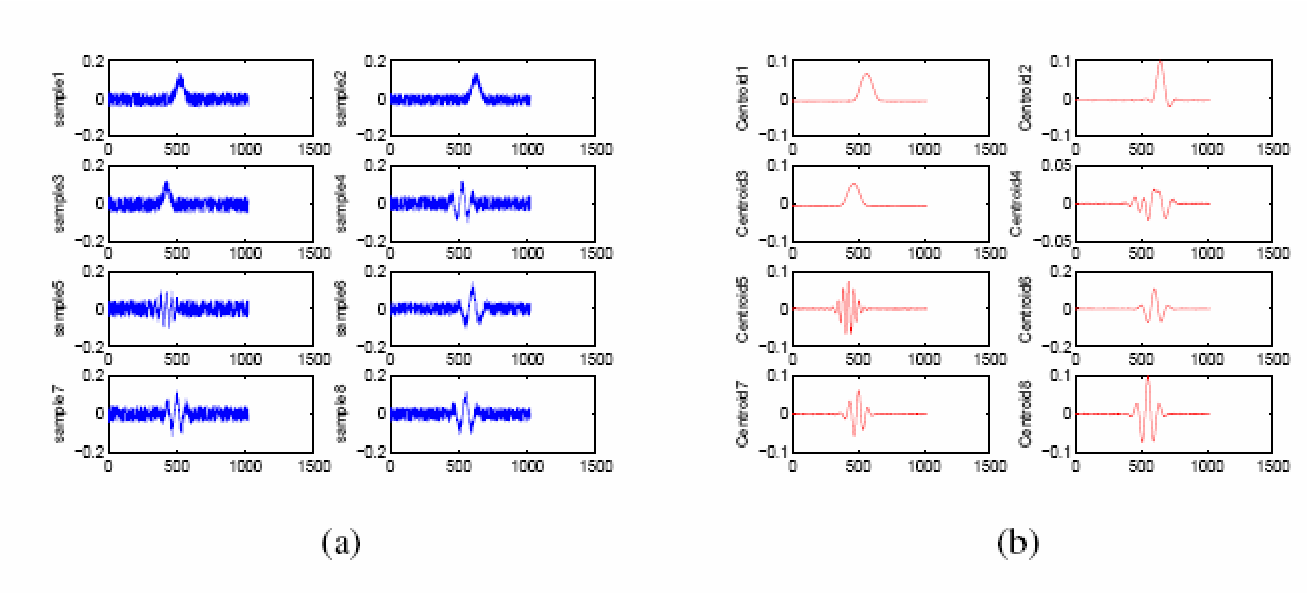


Fig. (a) Samples of simulated Gravitational-wave time series. (b) Centroids mined by S-means when *T = 0:1.*

Simulated GW trigger dataset: 20020 x 1024

S-means

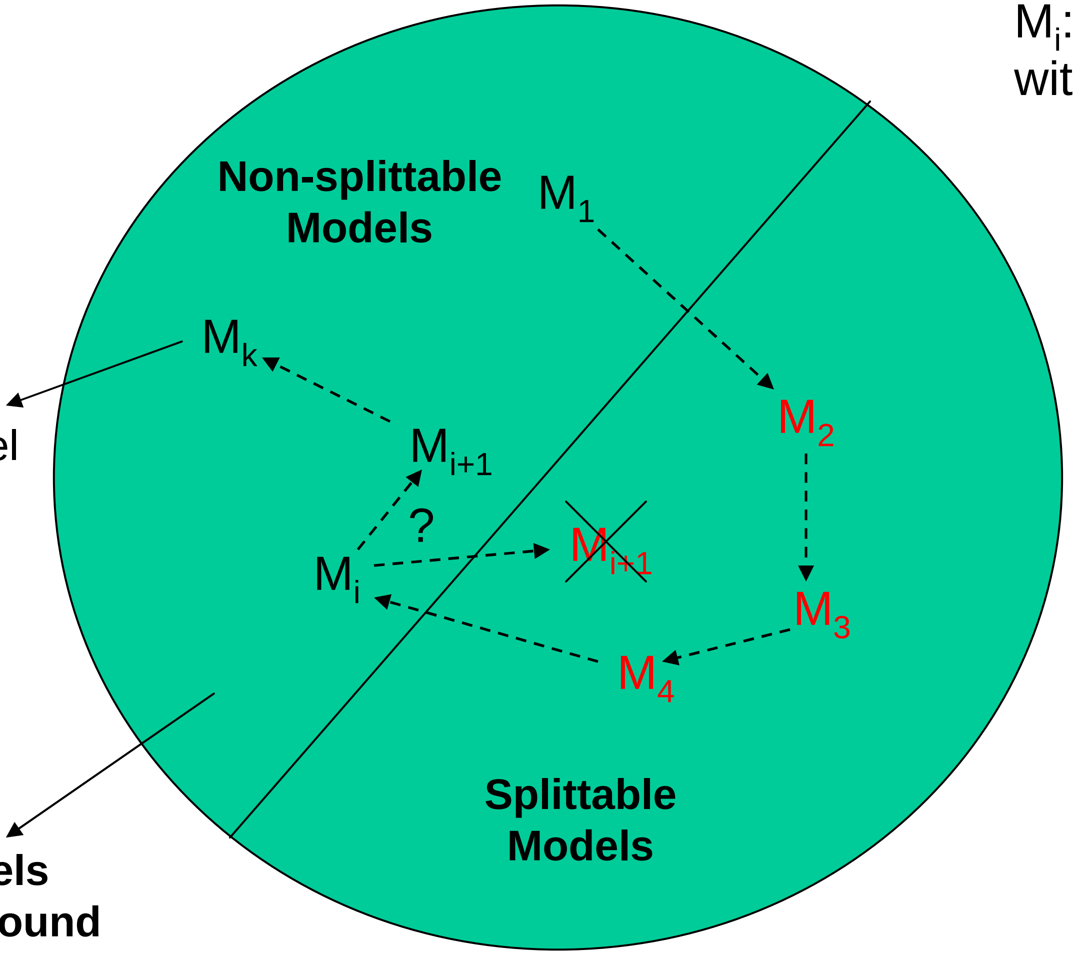
CONCLUSION AND FUTURE WORK

- S-means eliminates the necessity of specifying *K (the number of clusters) in K*-means clustering. An intuitive argument, similarity threshold *T is used instead of K* in S-means.
- *S-means mines the* number of compact clusters in a given dataset without prior knowledge in its statistical distribution.
- Extensive comparative experiments are needed to further validate the algorithm. For instance, the clustering result is very sensitive to threshold *T and the number of returned clusters can be unexpectedly* large when *T is high (e.g, $T > 0.4$).*
- *It is necessary to evaluate S-means* with different similarity measures such as Dynamic Time Warping and kernel functions in our future work.
- Estimated application to LIGO data – March 2008.

Constrained Validation Clustering Algorithm

Why Constraining the Search Space of Cluster Models?

M_i: a cluster model with i clusters



The correct cluster model

Correct models can only be found here (Note: this doesn’t mean that any model in this region is correct)

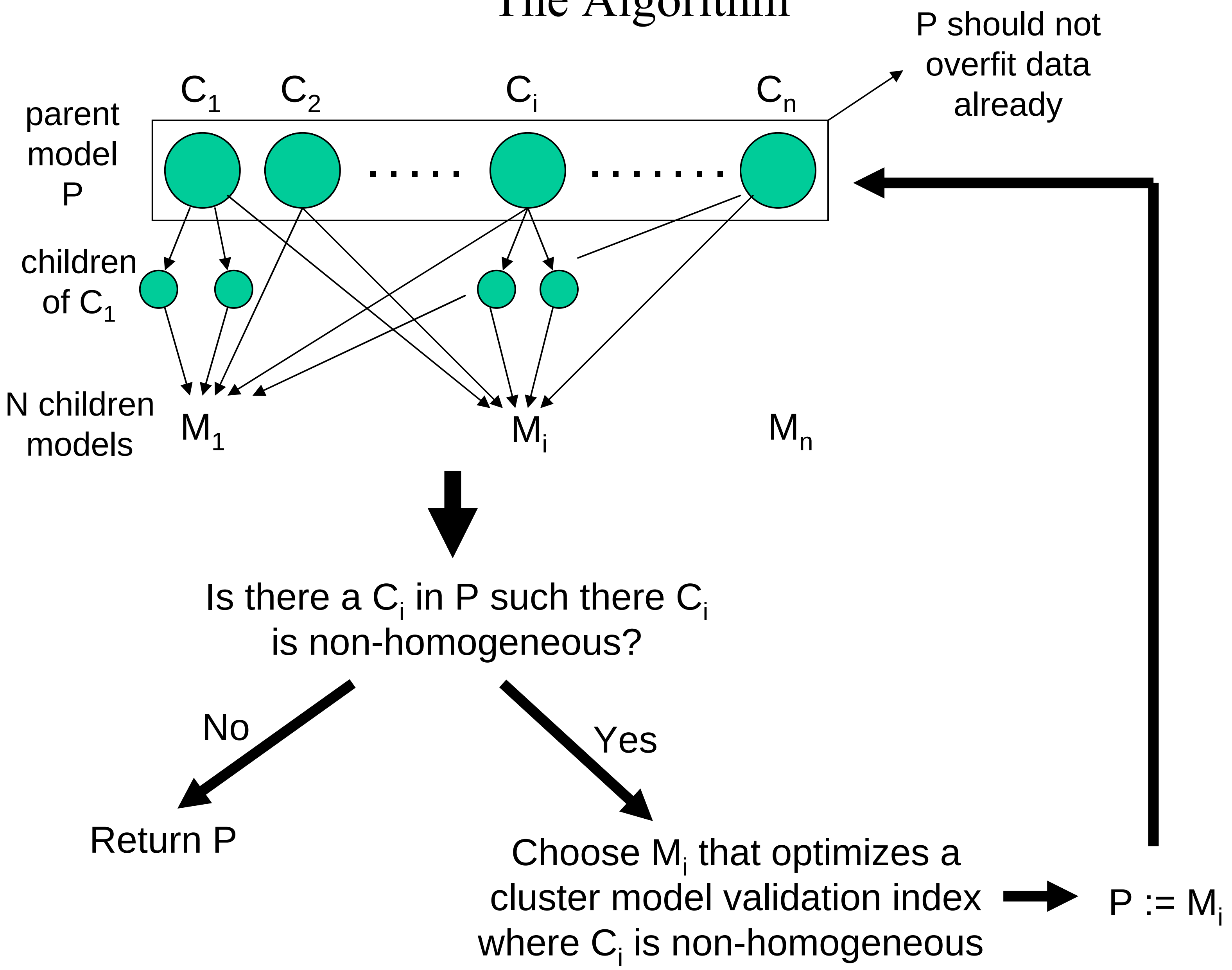
More Formal Outline of Key Ideas

- Suppose D is a set of data points in $\mathbf{R}^n$. The *source* of a subset of data points is a function *source*: $\mathcal{P}(D) \rightarrow \mathcal{P}(S)$ (mapping from powerset of D to powerset of S) where S is a set of stochastic processes (i.e. $S = \{F \mid F = \{F_t: t \in T\}\}$ where F_t is a random variable in a state space and T is "time").
- A *cluster model* $M = \{C_i \mid C_i$ is a subset of $D\}$is a partition of D.
- Suppose a cluster $C_i = X \cup Y$ such that $X \cap Y = \emptyset$. Then, X and Y are the *left child* and *right child* of C_i respectively.
- Given a cluster model M, two clusters C_i and C_j are *nearest* each other if the Euclidean distance between the centroids of C_i and C_j are shortest in M.
- A cluster C_i is *non-splittable* if the children X and Y are nearest to each other. Otherwise, C_i is *splittable*.
- A cluster model M is *non-splittable* if every cluster C_i in M is *non-splittable*.
- A cluster C is *homogeneous* if $|source(C)| = 1$. Otherwise, C is *non-homogeneous*.
- A set of data D has a *tolerable amount of noise* if it is the case that if $source(Y) = source(Z)$ and $source(Y) \neq source(X)$, then $d(mean(Y),mean(Z)) < d(mean(Y),mean(X))$ where X,Y,Z are subsets of D and $d(\ )$ is the Euclidean distance.

Theorem: Suppose D is a set of data with a tolerable amount of noise, and M_k is a cluster model in D with exactly k clusters. If for all cluster C_i in M_k, C_i is homogeneous and $source(C_i) \neq source(C_j)$ if $i \neq j$, then M_k is non-splittable.

Corollary: If cluster model M_k in D is splittable and D is a set of data with a tolerable amount of noise, then either there exists C_i in M_k that is non-homogeneous (i.e. M_k "underfits" the data) or it is the case that $source(C_i) = source(C_j)$ for some C_i and C_j where $i \neq j$ (i.e. M_k "overfits" the data).

The Algorithm



P should not overfit data already

Is there a C_i in P such there C_i is non-homogeneous?

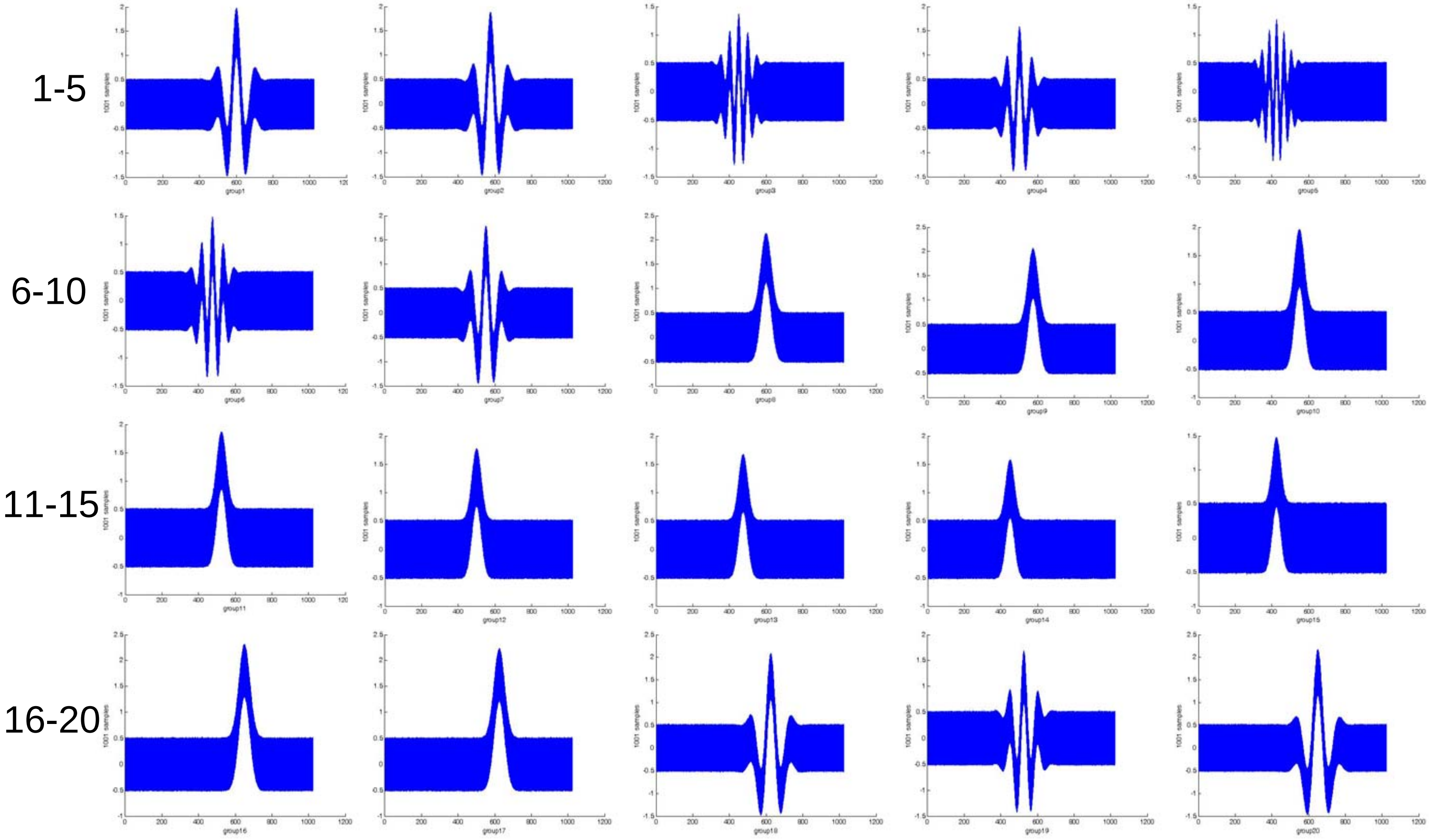
Return P

Choose M_i that optimizes a cluster model validation index where C_i is non-homogeneous

P := M_i

Experimental Results

Cluster



- Data size: 20020 x 1024 (20 clusters, each has a size of 1001)
- G-means found 2882 clusters (average size of a cluster = 6.9)

CV Clust conclusion and future direction

- G-means detected the existence of Gaussian clusters at the 95% confidence level – but unfortunately the model obviously over-fits the data.
- The model discovered by CV Clust was much more compact than that by G-means.
- A future work is to provide statistical meaning to results discovered by CV Clust
- Tentative schedule of implementation completion – March 2008
- Tentative schedule of application to LIGO data – Fall 2008