An Introduction to Model-Based Clustering

Anish R. Shah, CFA
Northfield Information Services
Anish@northinfo.com

Newport
June 3, 2011

Northfield
INFORMATION SERVICES, INC.

Clustering

- Observe characteristics of some objects
  - $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ N objects
- Goal: group alike objects
  - say there are M clusters
    $\{z_1, \ldots, z_N\}$ cluster memberships
    z_k = object k's membership, a number from 1..M
  - k, j in the same cluster → $\mathbf{x}_k$, $\mathbf{x}_j$ similar
    -or- k, j in different clusters → $\mathbf{x}_k$, $\mathbf{x}_j$ dissimilar

Northfield
INFORMATION SERVICES, INC.

Examples of Characteristics

- Clustering dog breeds
  - **x** = (snout length / width of face,
    dog's BMI
    fur type [*wiry, short, long, none*])
- Clustering stocks via returns
  - **x** = (past 3 years of monthly returns)
- Clustering stocks via fundamentals
  - **x** = (beta to the market, dividend rate, E/P, Debt/Equity, …)
- Clustering stocks via fundamentals & returns
  - **x** = (beta to the market, dividend rate,
    past 2 years of monthly returns)

Northfield
INFORMATION SERVICES, INC.

Machine Learning

- Rather than being programmed with rules, the system inferentially learns the patterns/rules of reality from data
- Supervised Learning
  - Some of the training data is labeled
  - e.g. There are 5 company types - AAPL & MSFT are type 1, ... , XOM is type 5. Find the prototype for each type and label the rest of the universe
  - e.g. Amazon & Netflix recommendations
- Unsupervised Learning
  - None of the data has labels
  - Organize the system to maximize some criterion
  - e.g. Clustering maximizes similarity within each cluster
  - e.g. Principal Components Analysis maximizes explained variance
  - Vanilla clustering is the canonical example of unsupervised machine learning

Northfield
INFORMATION SERVICES, INC.

Review of Forms of Hard Clustering

- ‘Hard’ means an object is assigned to only one cluster
  - In contrast, model-based clustering can give a probability distribution over the clusters

- Hierarchical Clustering
  - Maximize distance between clusters
  - Flavors come from different ways of measuring distance
    - Single Linkage – distance between the two nearest elements
    - Complete Linkage – distance between the two farthest elements
    - Average Linkage – mean (or median) distance between all elements

- K-Means
  - Minimize mean (median in K-medians) distance within clusters

K-Means / K-Medians

- K-Means (heuristically) assigns objects to clusters to minimize the average squared distance (absolute distance in K-Medians) from object to cluster center.

- Minimize $\frac{1}{N} \sum_{k=1..N} ||\mathbf{x}_k - \boldsymbol{\mu}_{z_k}||^2$
  over

  $z_1..z_N$ = cluster assignments

  $\boldsymbol{\mu}_1..\boldsymbol{\mu}_M$ = centers of the clusters

Northfield
INFORMATION SERVICES, INC.

K-Means Algorithm

1. Randomly assign objects to clusters

2. Calculate the center (mean) of each cluster
3. Check assignments for all the objects. If another center is closer to an object, reassign the object to that cluster

4. Repeat steps 2-3 until no reassignments occur

- Extremely fast
- The solution is a local max, so several starting points are used in practice
- (K-Medians) For robustness, step 2 uses median instead of mean to get centers

Northfield
INFORMATION SERVICES, INC.

Mixture of Gaussians: A Model-Based Clustering Similar to K-Means

- Observe data for N objects, $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$
- Each cluster generates data distributed normally around its center
  - when object k is from cluster m,
    $p(\mathbf{x}_k) \sim \exp(\| \mathbf{x}_k - \boldsymbol{\mu}_m \|^2 / \sigma^2)$
- Some clusters appear more frequently than others
  - given no observation information,
    p(an object belongs to cluster m) = π_M
- Find the setup that make the observations most likely to occur
  - cluster centers $\{\boldsymbol{\mu}_1 .. \boldsymbol{\mu}_M\}$
  - variance σ^2
  - cluster frequencies $\{\pi_1 .. \pi_M\}$

Northfield
INFORMATION SERVICES, INC.

Model-Based Clustering

- Observe characteristics of some objects
  - $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ N objects
- An object belongs to one of M clusters, but you don't know which
  - $\{z_1, \ldots, z_N\}$ cluster memberships, numbers from 1..M
- Some clusters are more likely than others
  - $P(z_k=m) = \pi_m$ (π_m = frequency cluster m occurs)
- Within a cluster, objects' characteristics are generated by the same distribution, which has free parameters
  - $P(\mathbf{x}_k | z_k=m) = f(\mathbf{x}_k, \boldsymbol{\lambda}_m)$ ($\boldsymbol{\lambda}_m$ = parameters of cluster m)
  - f doesn't have to be Gaussian

Northfield INFORMATION SERVICES, INC.

Model-Based Clustering (2)

- Now you have a model connecting the observations to the cluster memberships and parameters
  - $P(\mathbf{x}_k) = \sum_{m=1..M} P(\mathbf{x}_k | z_k{=}m)\, P(z_k{=}m)$
  - $\quad = \sum_{m=1..M} f(\mathbf{x}_k, \boldsymbol{\lambda}_m)\, \pi_m$
  - $P(\mathbf{x}_1 \ldots \mathbf{x}_N) = \prod_{k=1..N} P(\mathbf{x}_k)$ (assuming **x**'s are independent)

1. Find the values of the parameters by maximizing the likelihood (usually the log of the likelihood) of the observations
   - max log $P(\mathbf{x}_1 \ldots \mathbf{x}_N)$ over $\boldsymbol{\lambda}_1 \ldots \boldsymbol{\lambda}_M$ and $\pi_1 \ldots \pi_M$
   - This turns out to be a nonlinear mess and is greatly aided by the "EM Algorithm" (next slide)

2. With parameters in hand, calculate the probability of membership given the observations
   - $P(z|\mathbf{x}) = P(\mathbf{x}|z)\, P(z) / P(\mathbf{x})$

Northfield
INFORMATION SERVICES, INC.

EM (Expectation-Maximization) Algorithm Setup

- Let $\theta = (\boldsymbol{\lambda}_1 \ldots \boldsymbol{\lambda}_M, \pi_1 \ldots \pi_M)$, the parameters being maximized over
- Observe x. Don't know z, the cluster memberships
- Want to maximize $\log p(x|\theta)$, but it is too complicated
- EM can be used when
  - It's possible to make an approximation of $p(z|x,\theta)$, the conditional distribution of the hidden variables
  - $\log p(x,z|\theta)$, the probability if all the variables were known, is easy to manipulate

Northfield INFORMATION SERVICES, INC.

The EM Algorithm

- Want to maximize $\log p(x|\theta) = \log \int p(x,z|\theta)\, dz$
- (E Step)
  - Create an approximate distribution of the missing data. Call it $u(z)$
    Ideally this is $p(z|x,\theta)$
  - Let $Q(\theta)$ = the log likelihood under θ averaged by $u(z)$
    $= \int \log p(x,z|\theta)\, u(z)\, dz$
- (M Step)
  - Maximize $Q(\theta)$ over θ
  - θ_{new} = the maximizer

- Repeat E & M steps until convergence

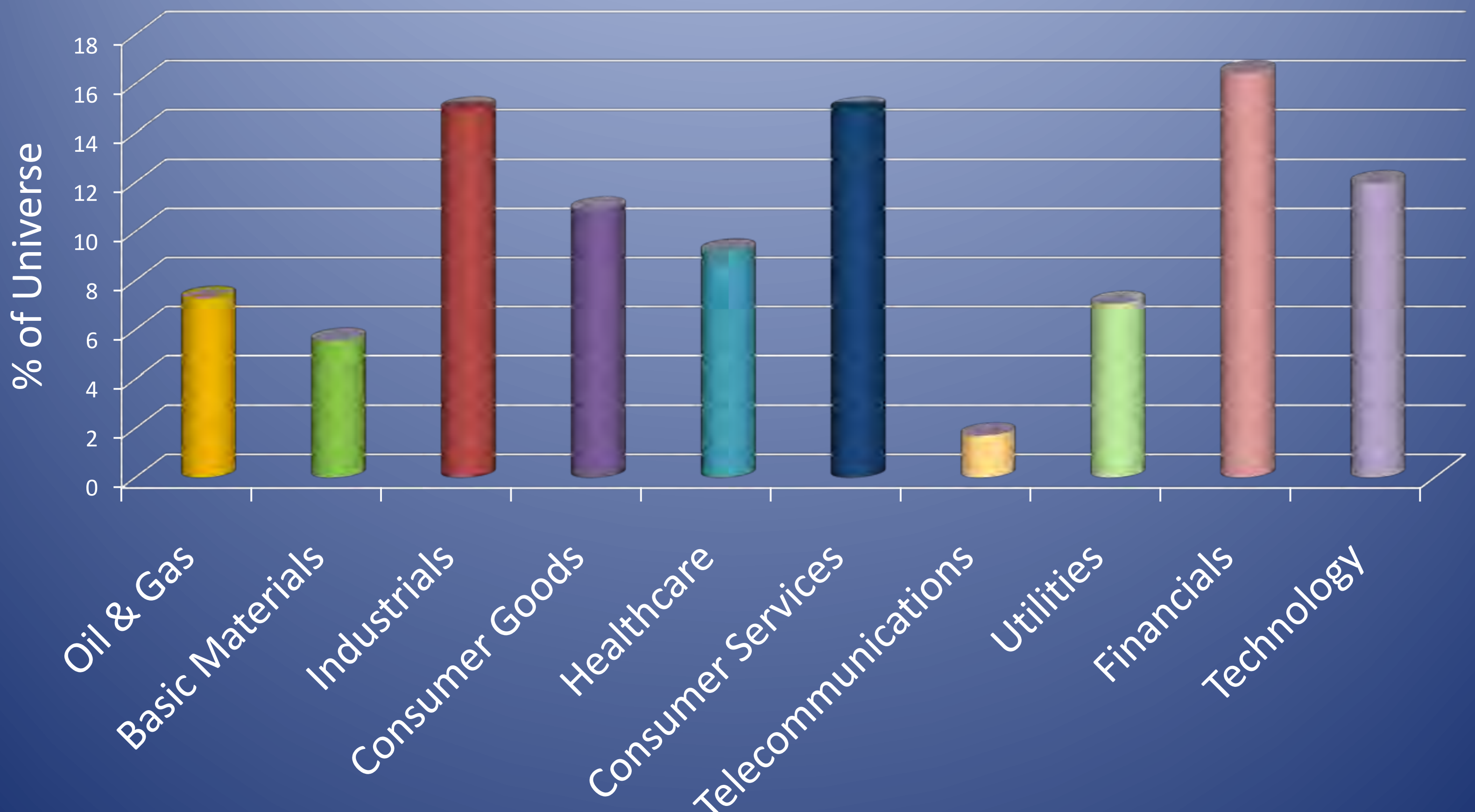
- EM switches between 1) finding an approximate distribution of missing data given the parameters and 2) finding better parameters given the approximation

Northfield INFORMATION SERVICES, INC.

Experiments

- Universe is the "S&P 468" – the S&P 500 stripped of securities missing data
- Know ICB sector assignments for these companies
  - 10 sectors: Oil & Gas, Basic Materials, Industrials, Consumer Goods, Healthcare, Consumer Services, Telecommunications, Utilities, Financials, Technology
- Have information about the companies
  - 5 years of monthly returns
  - market β
  - fundamentals (E/P, B/P, rev/P, debt/equity, yield, trading activity, relative strength, log mkt cap, earnings variability, growth rate, price volatility)
  - Each characteristic is scaled to make its cross-sectional standard deviation 1
- Using assortments of information, cluster securities into 10 groups

Universe Breakdown by Sector



Cast Clusters in Terms of the Known Sectors

- To illustrate results in this presentation, cast the clusters in terms of the original sectors
  - For each sector, sum the cluster probabilities of all companies in that sector. Rescale so the sum is 1. This gives sectors in terms of clusters
  - Take that rescale in the other direction, so each cluster sums to 1. This gives the clusters in terms of sectors, without biasing toward numerous sectors

- There are many other uses for the cluster results

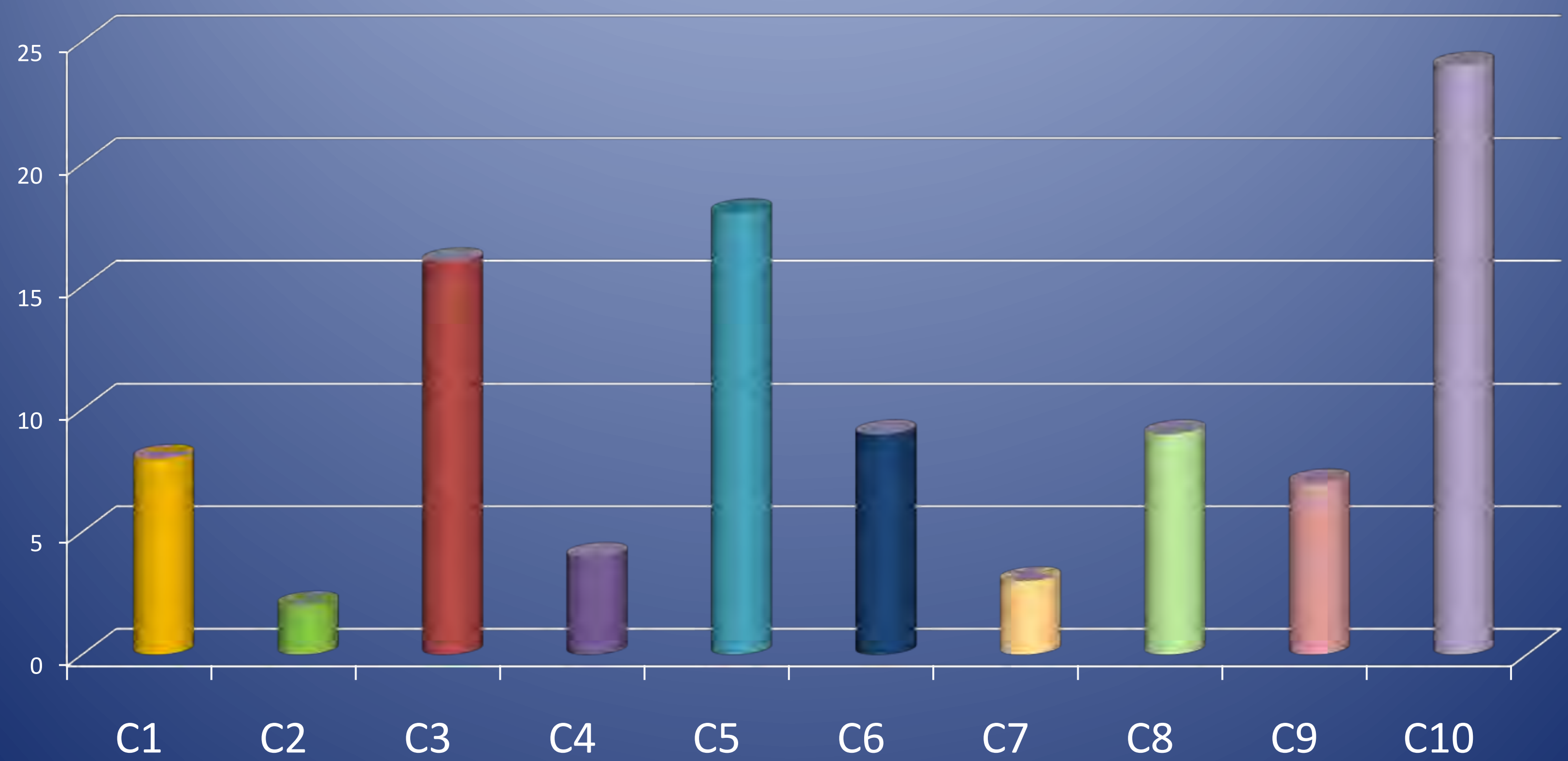
Northfield
INFORMATION SERVICES, INC.

Experiment 1:

5 Years Monthly Returns

Gaussian Mixture Model

% of Probability Mass by Cluster

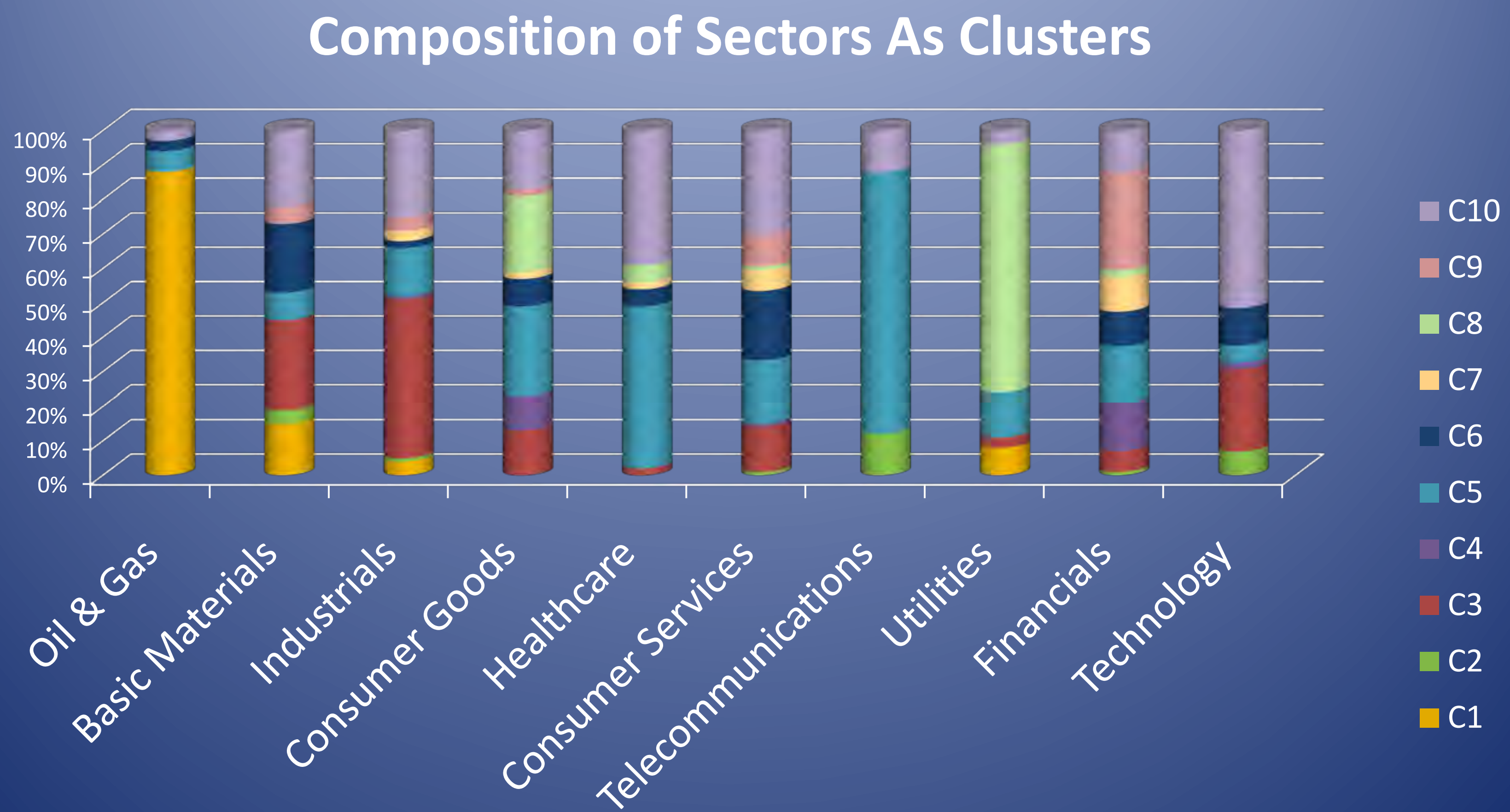


Northfield
INFORMATION SERVICES, INC.

Experiment 1:

5 Years Monthly Returns

Gaussian Mixture Model

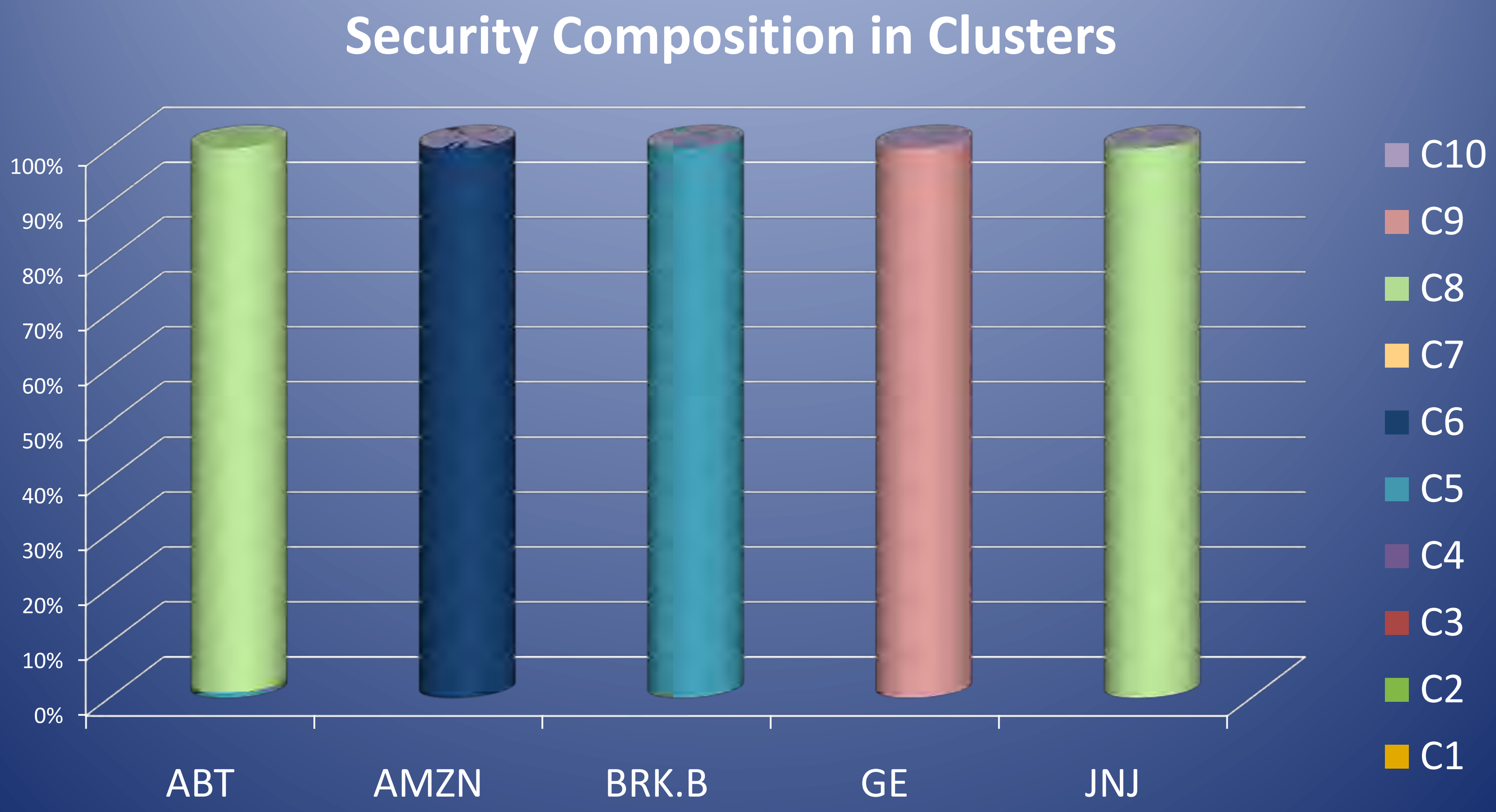


Northfield
INFORMATION SERVICES, INC.

Experiment 1:

5 Years Monthly Returns

Gaussian Mixture Model



Northfield INFORMATION SERVICES, INC.

Experiment 1:

5 Years Monthly Returns

Gaussian Mixture Model

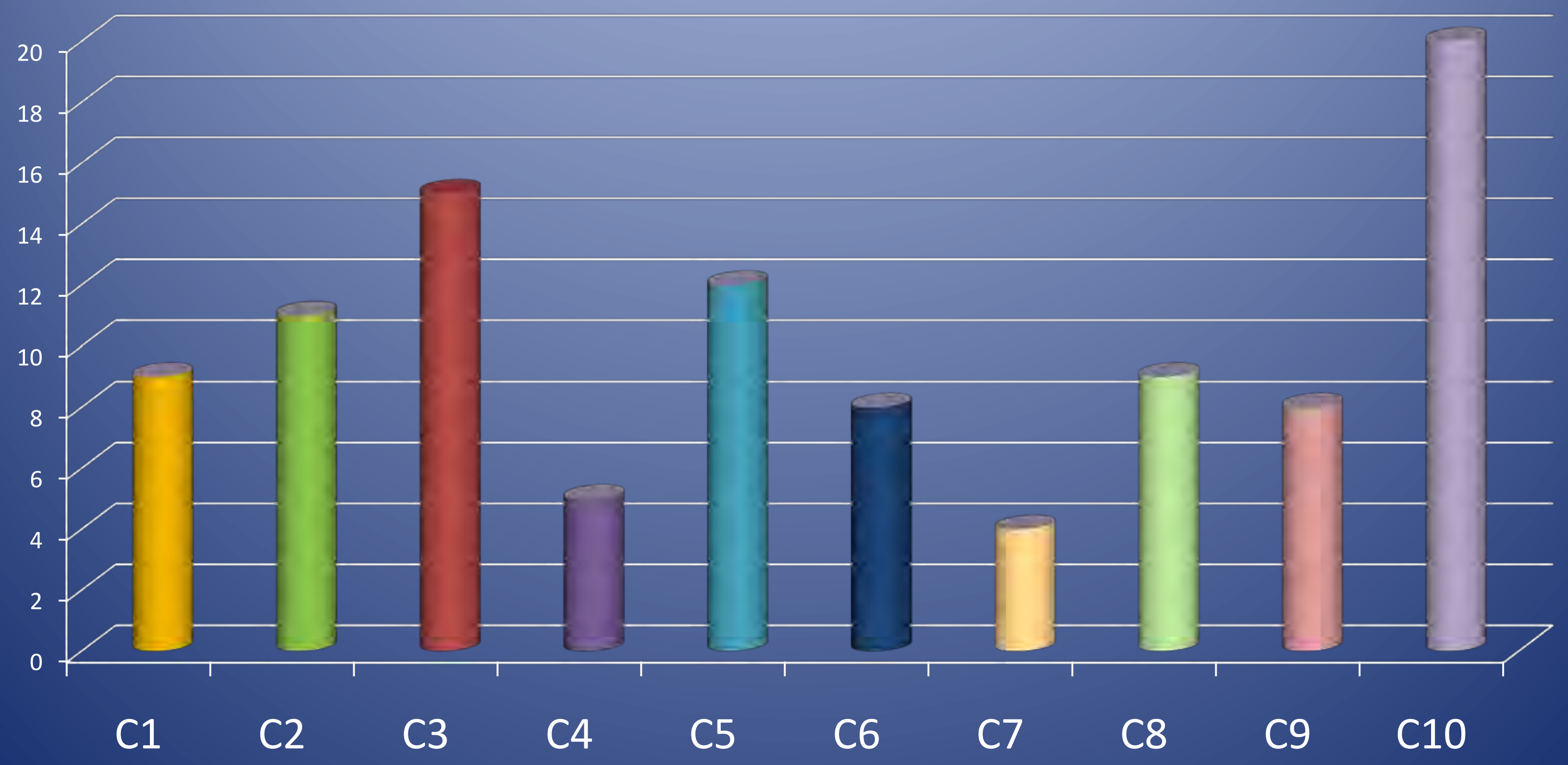
Northfield INFORMATION SERVICES, INC.

Experiment 2:

5 Years Monthly Returns & β

Gaussian Mixture Model

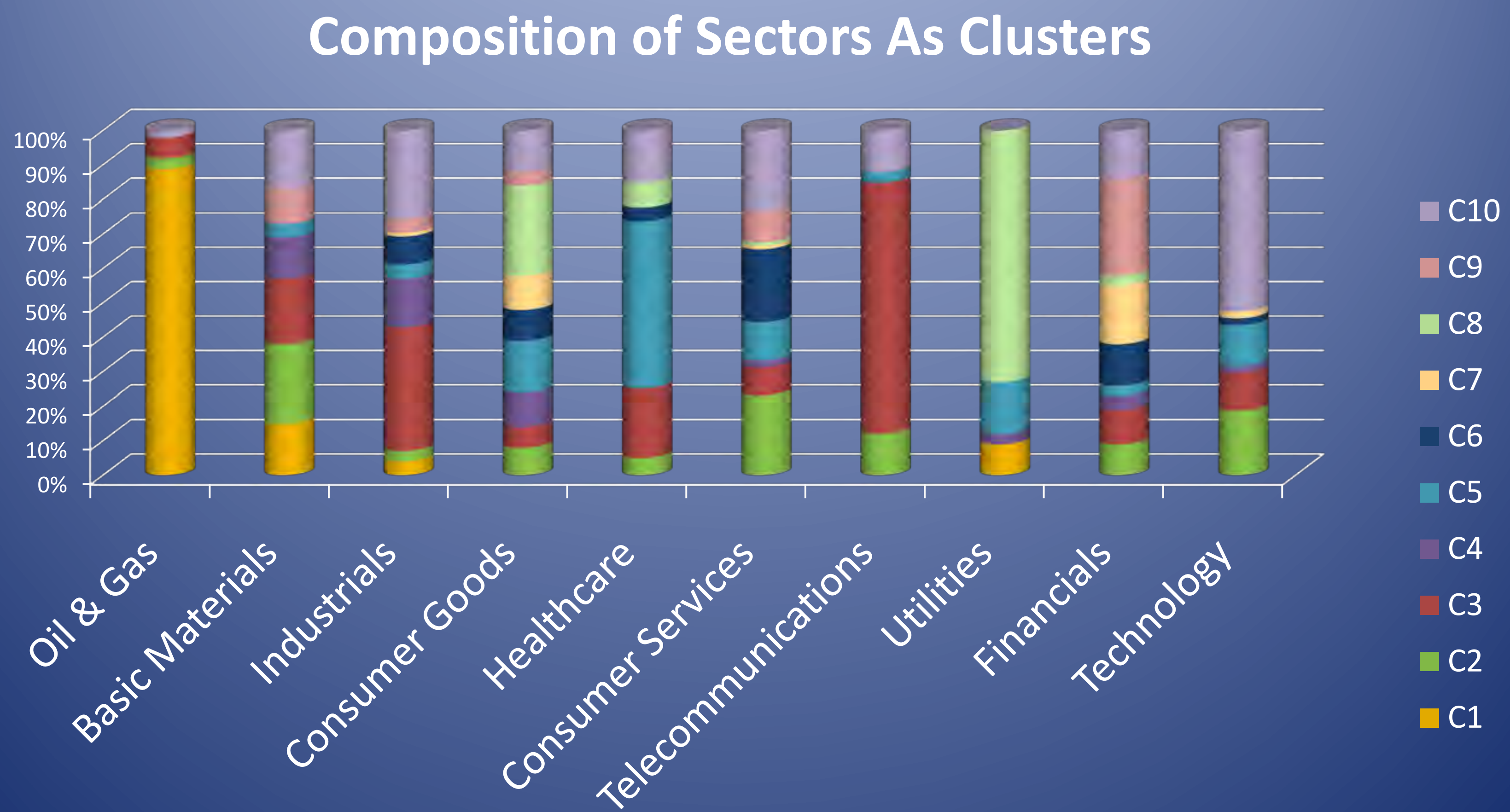
% of Probability Mass by Cluster



Experiment 2:

5 Years Monthly Returns & β

Gaussian Mixture Model

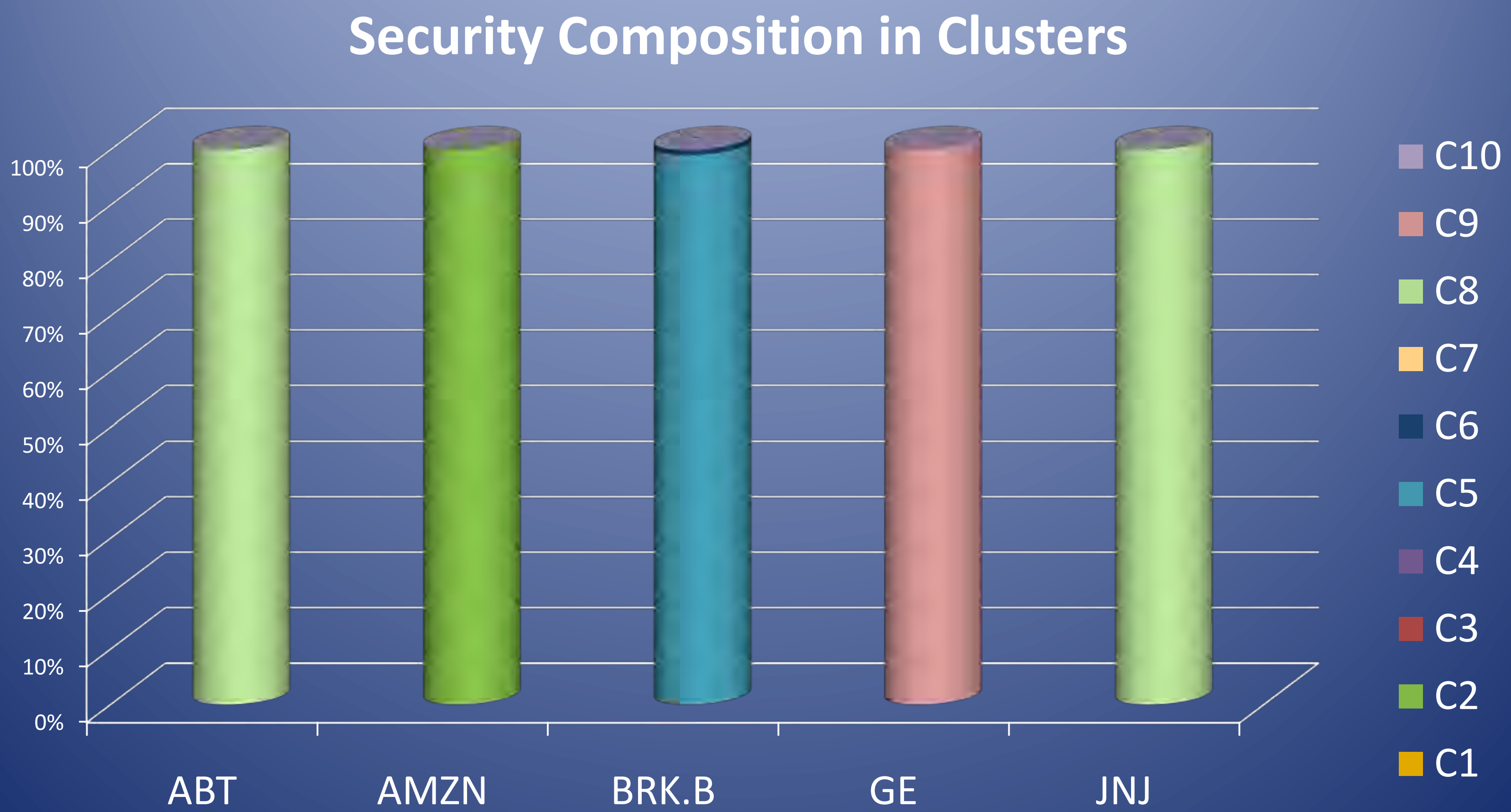


Northfield
INFORMATION SERVICES, INC.

Experiment 2:

5 Years Monthly Returns & β

Gaussian Mixture Model



Northfield
INFORMATION SERVICES, INC.

Experiment 2:

5 Years Monthly Returns & β

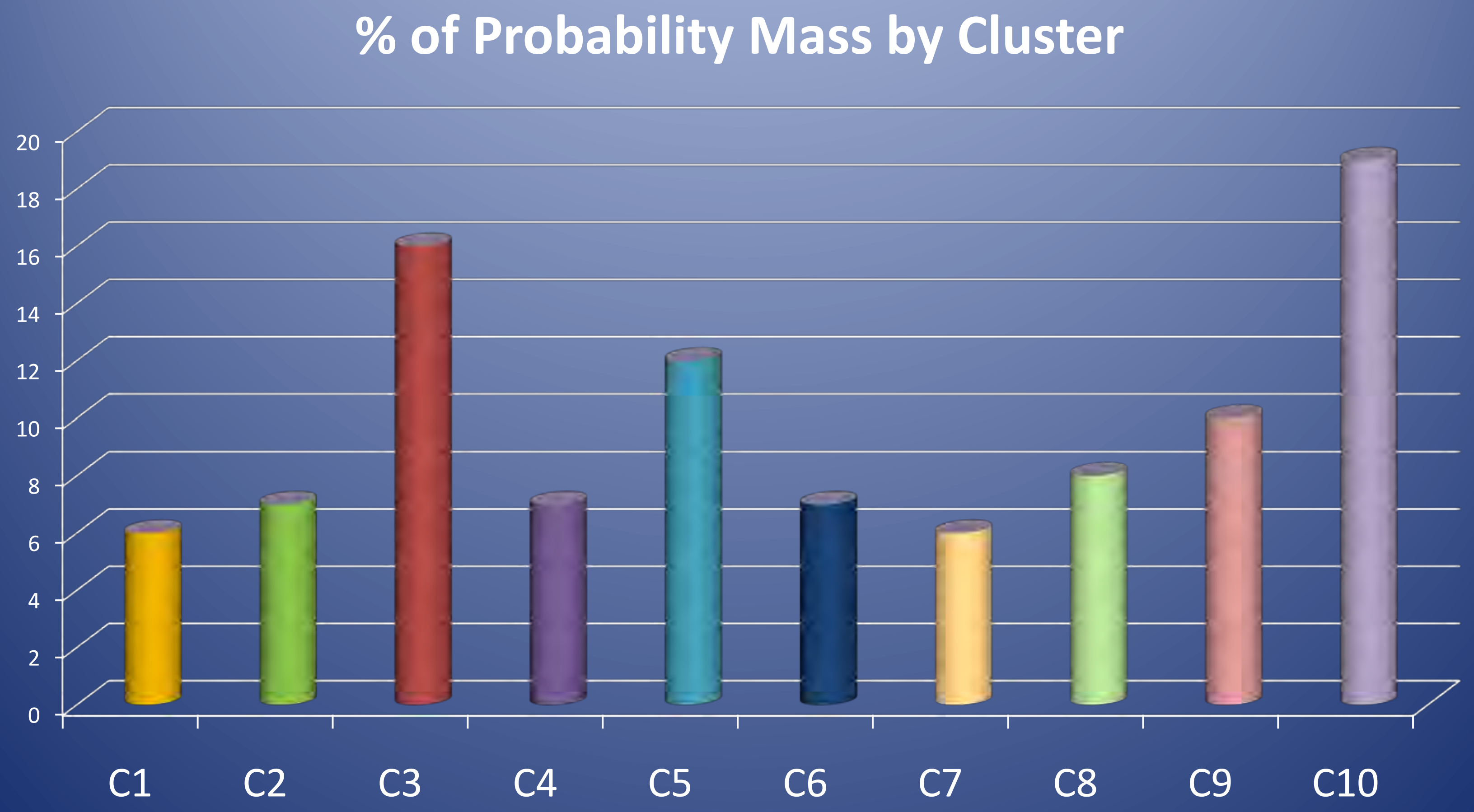
Gaussian Mixture Model

Northfield
INFORMATION SERVICES, INC.

Experiment 3:

Fundamentals – E/P, E/B, …

Gaussian Mixture Model

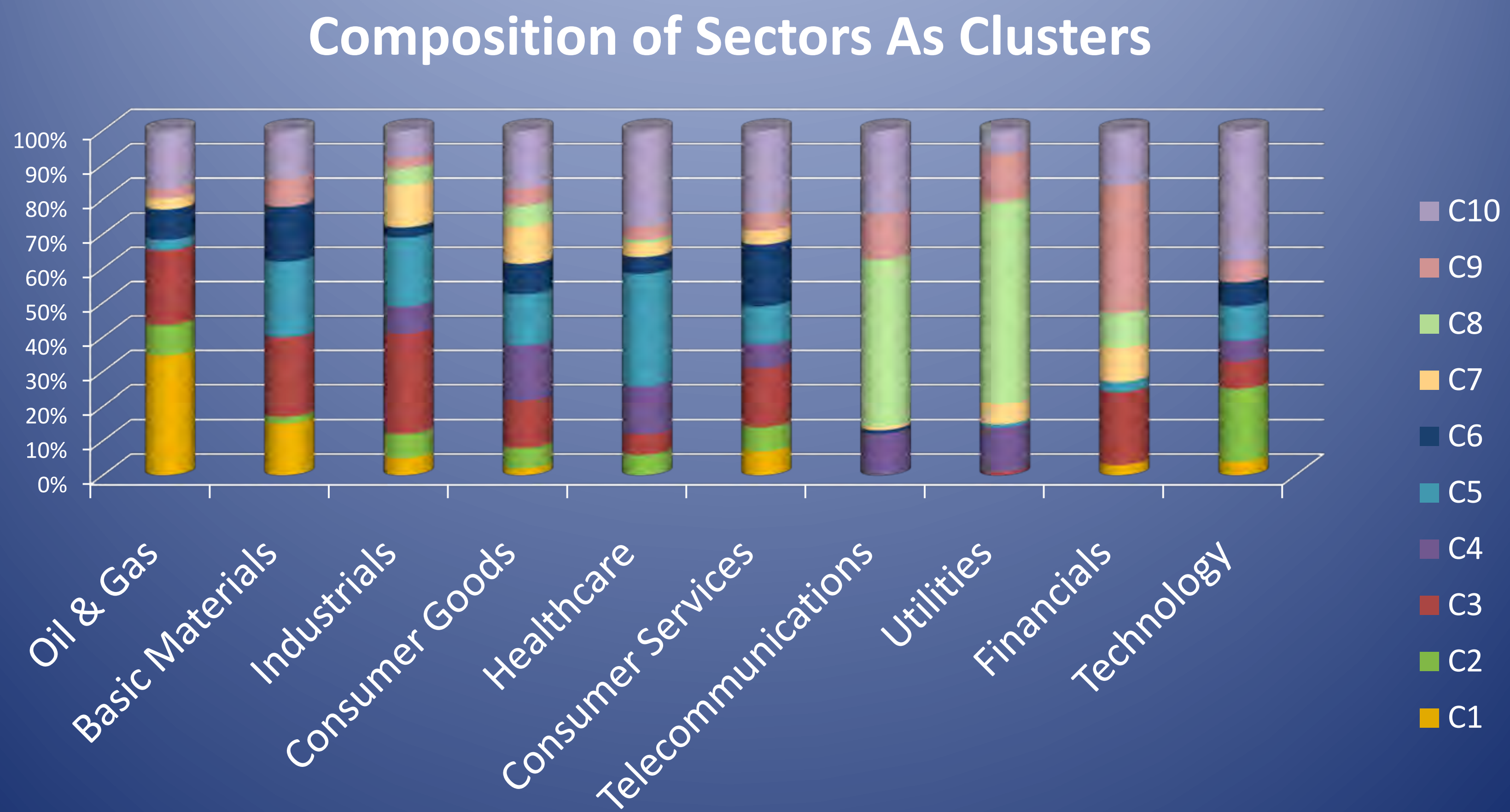


Northfield
INFORMATION SERVICES, INC.

Experiment 3:

Fundamentals – E/P, E/B, …

Gaussian Mixture Model

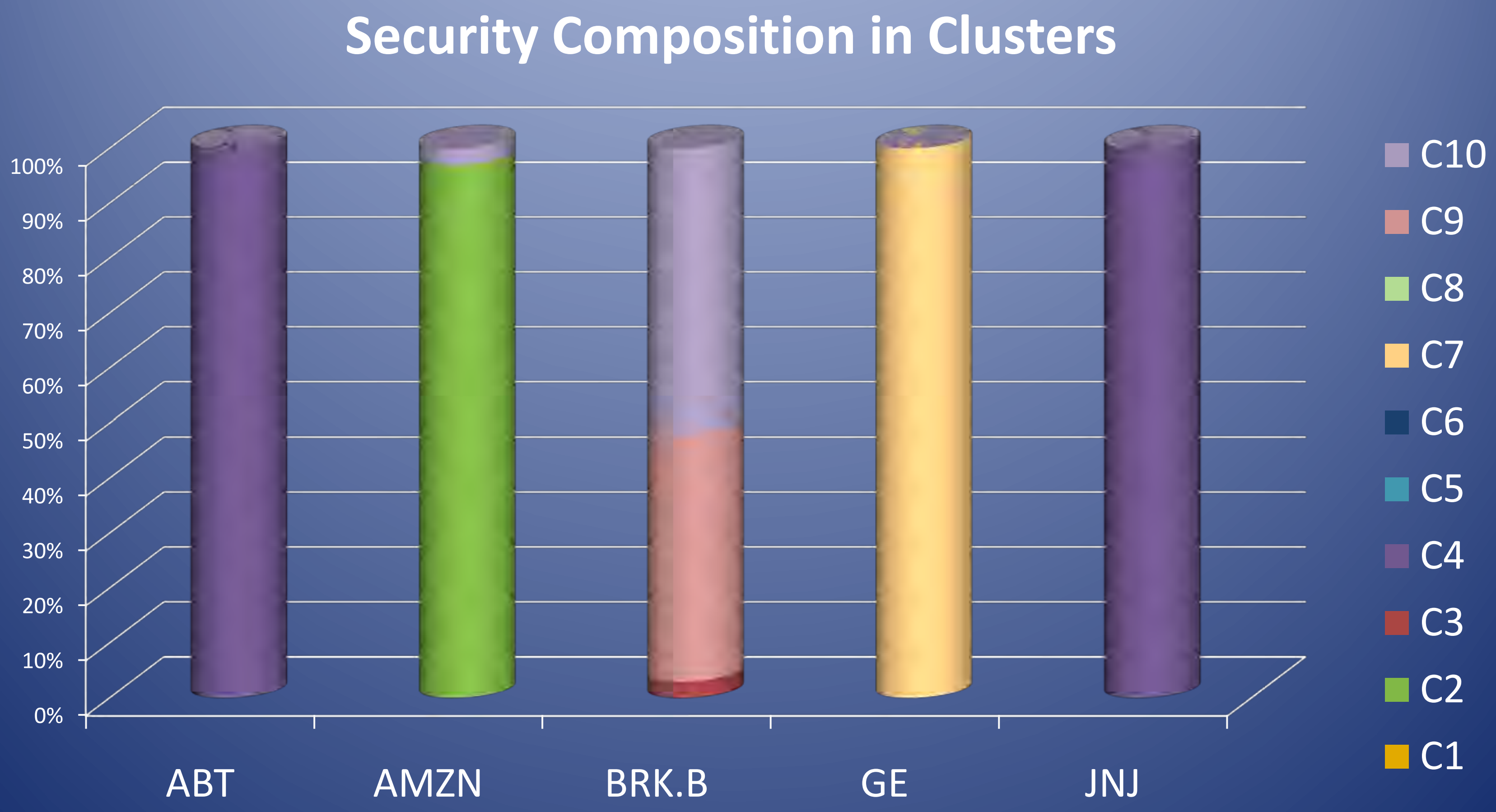


Northfield
INFORMATION SERVICES, INC.

Experiment 3:

Fundamentals – E/P, E/B, …

Gaussian Mixture Model



Northfield
INFORMATION SERVICES, INC.

Experiment 3:

Fundamentals – E/P, E/B, …

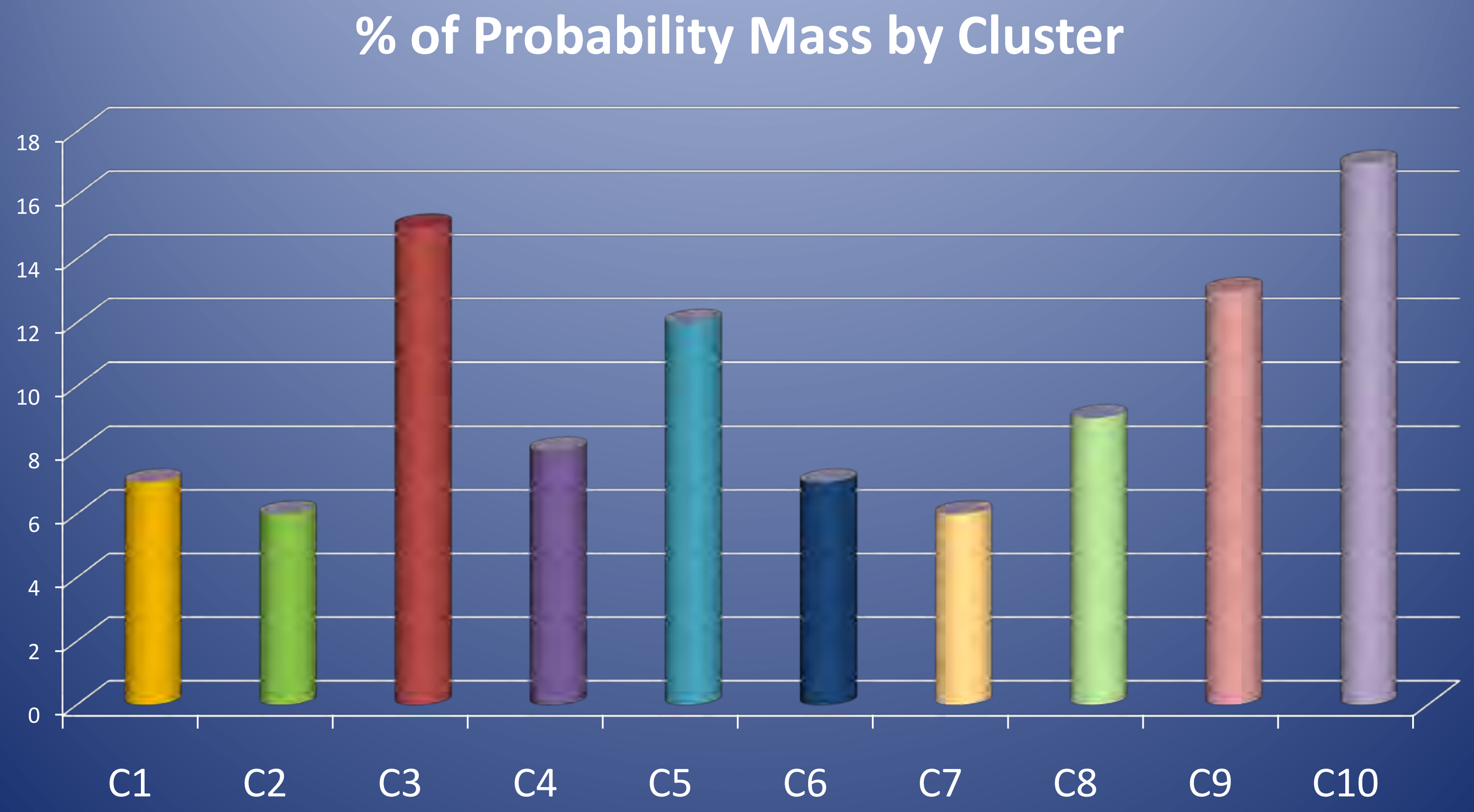
Gaussian Mixture Model

Northfield
INFORMATION SERVICES, INC.

Experiment 4:

Fundamentals & β

Gaussian Mixture Model

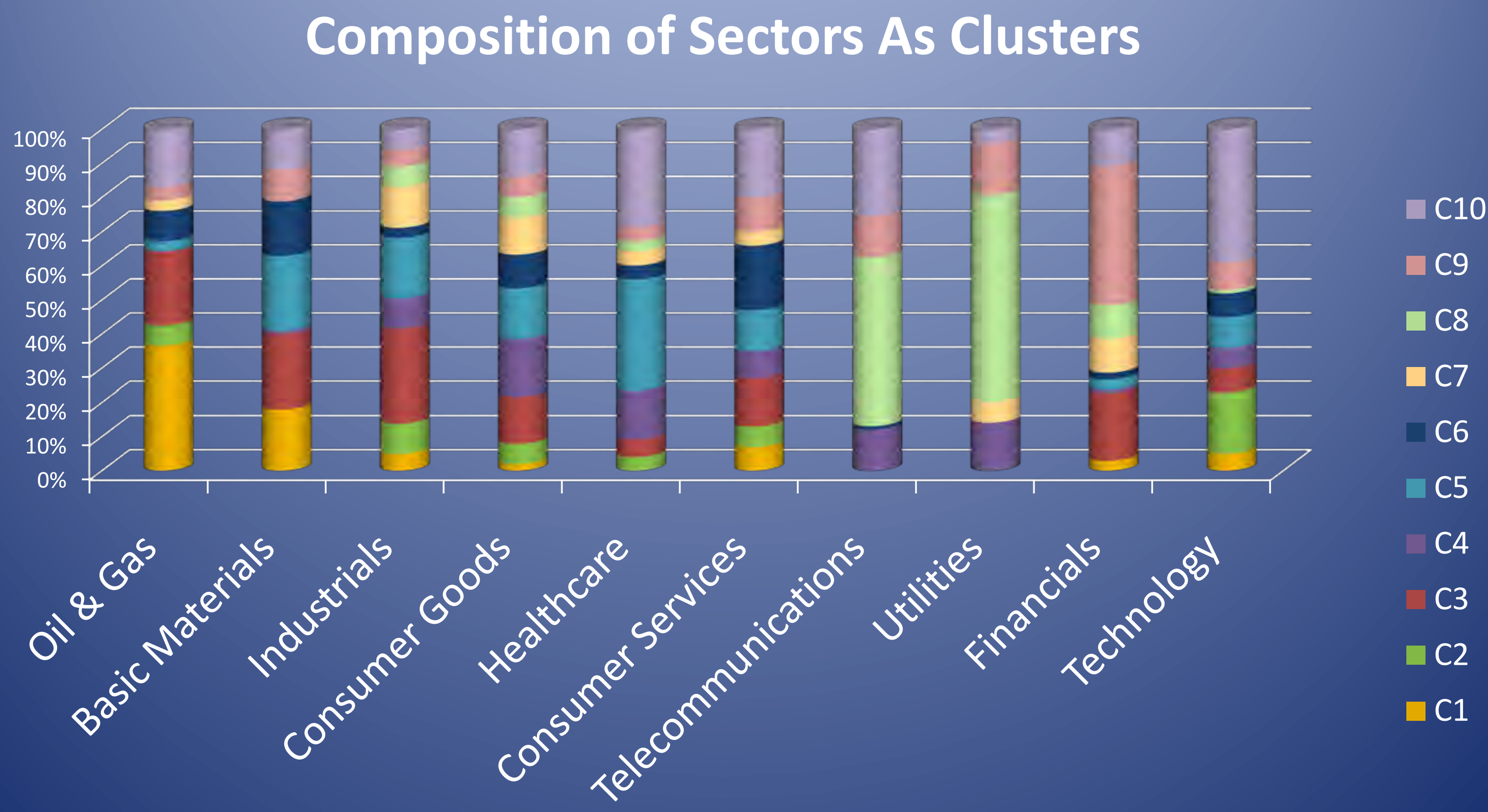


Northfield
INFORMATION SERVICES, INC.

Experiment 4:

Fundamentals & β

Gaussian Mixture Model

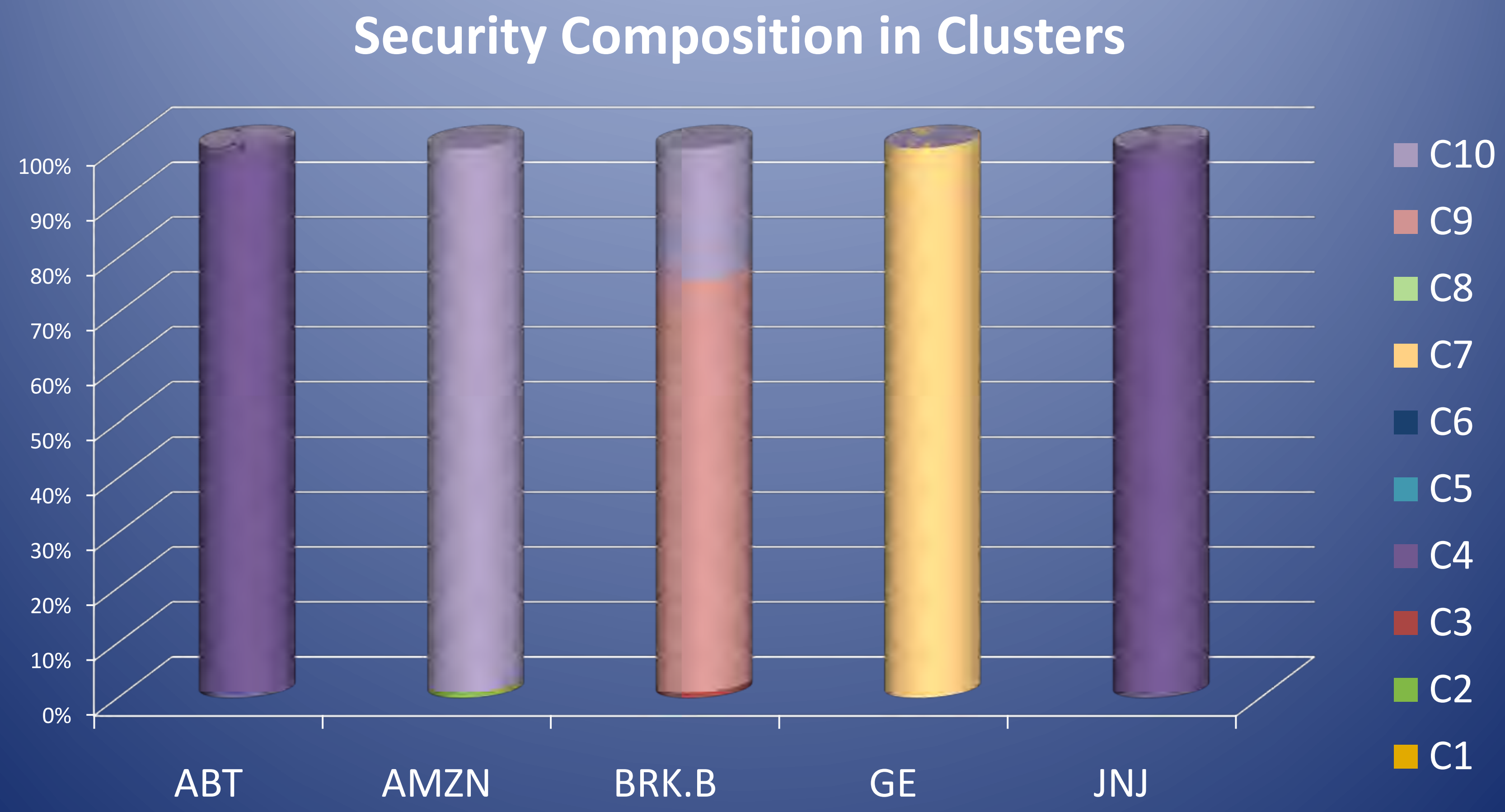


Northfield INFORMATION SERVICES, INC.

Experiment 4:

Fundamentals & β

Gaussian Mixture Model



Northfield
INFORMATION SERVICES, INC.

Experiment 4:

Fundamentals & β

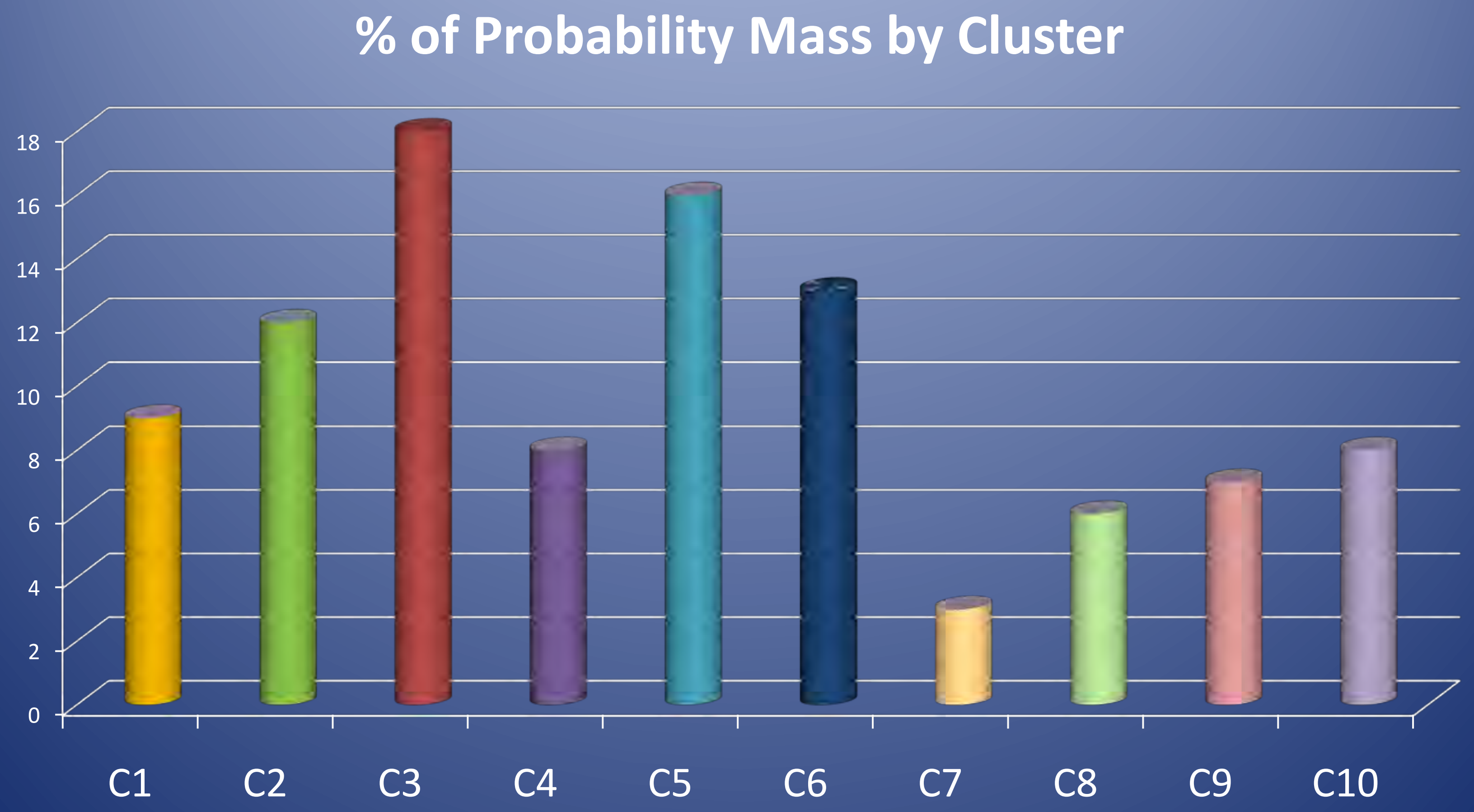
Gaussian Mixture Model

Northfield
INFORMATION SERVICES, INC.

Experiment 5:

5 Years Returns, Fundamentals & β

Gaussian Mixture Model

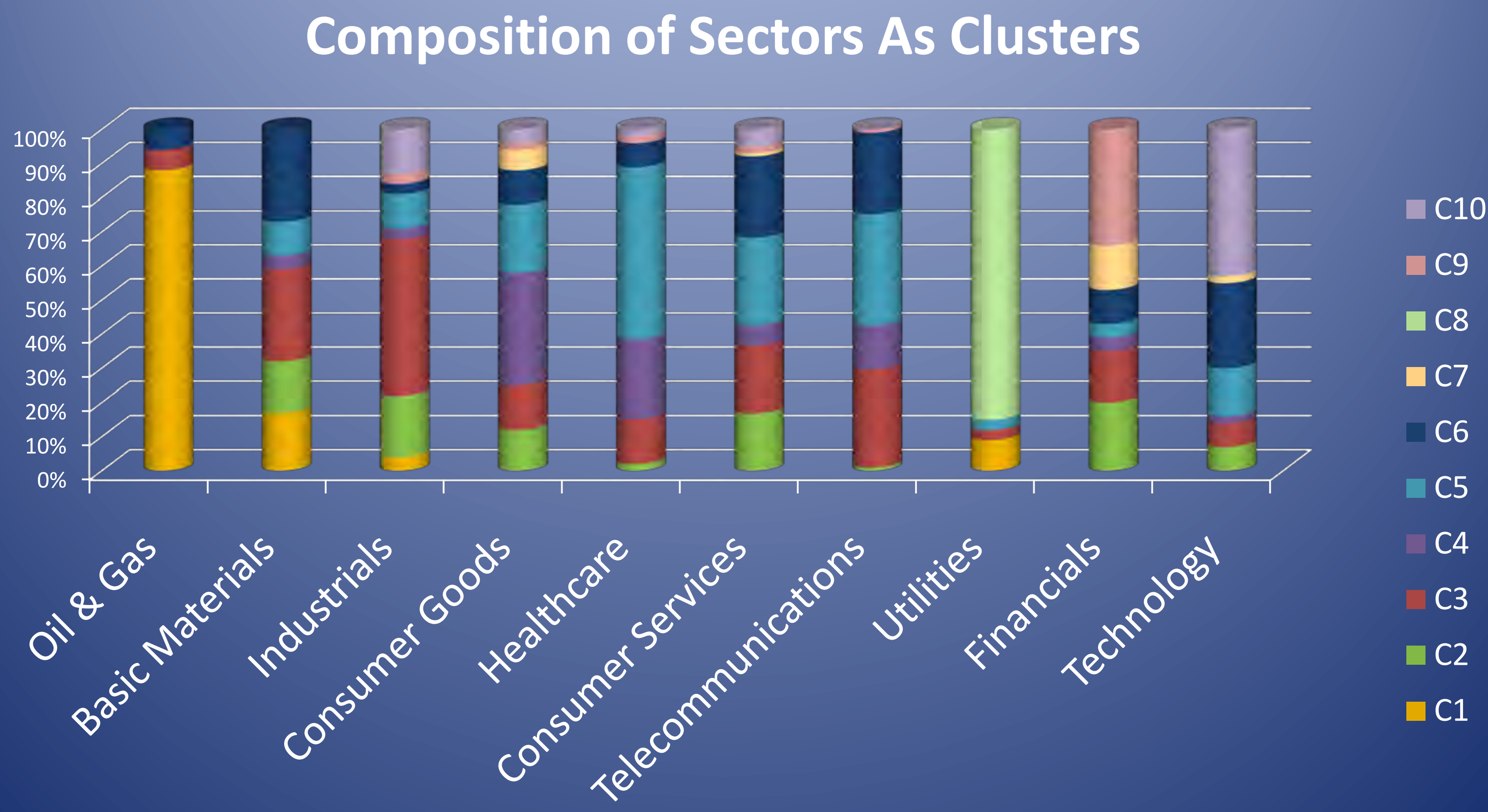


Northfield
INFORMATION SERVICES, INC.

Experiment 5:

5 Years Returns, Fundamentals & β

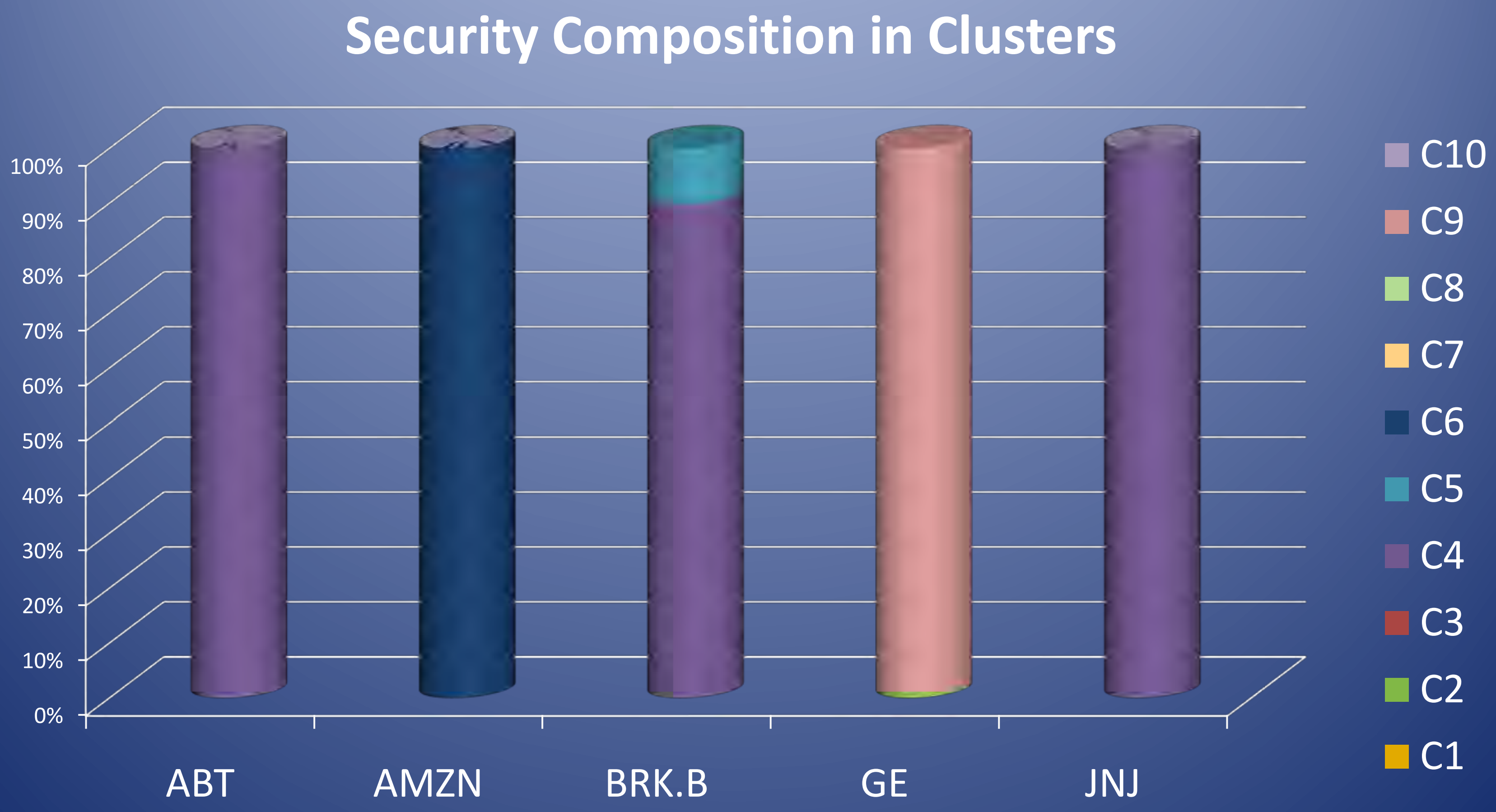
Gaussian Mixture Model



Experiment 5:

5 Years Returns, Fundamentals & β

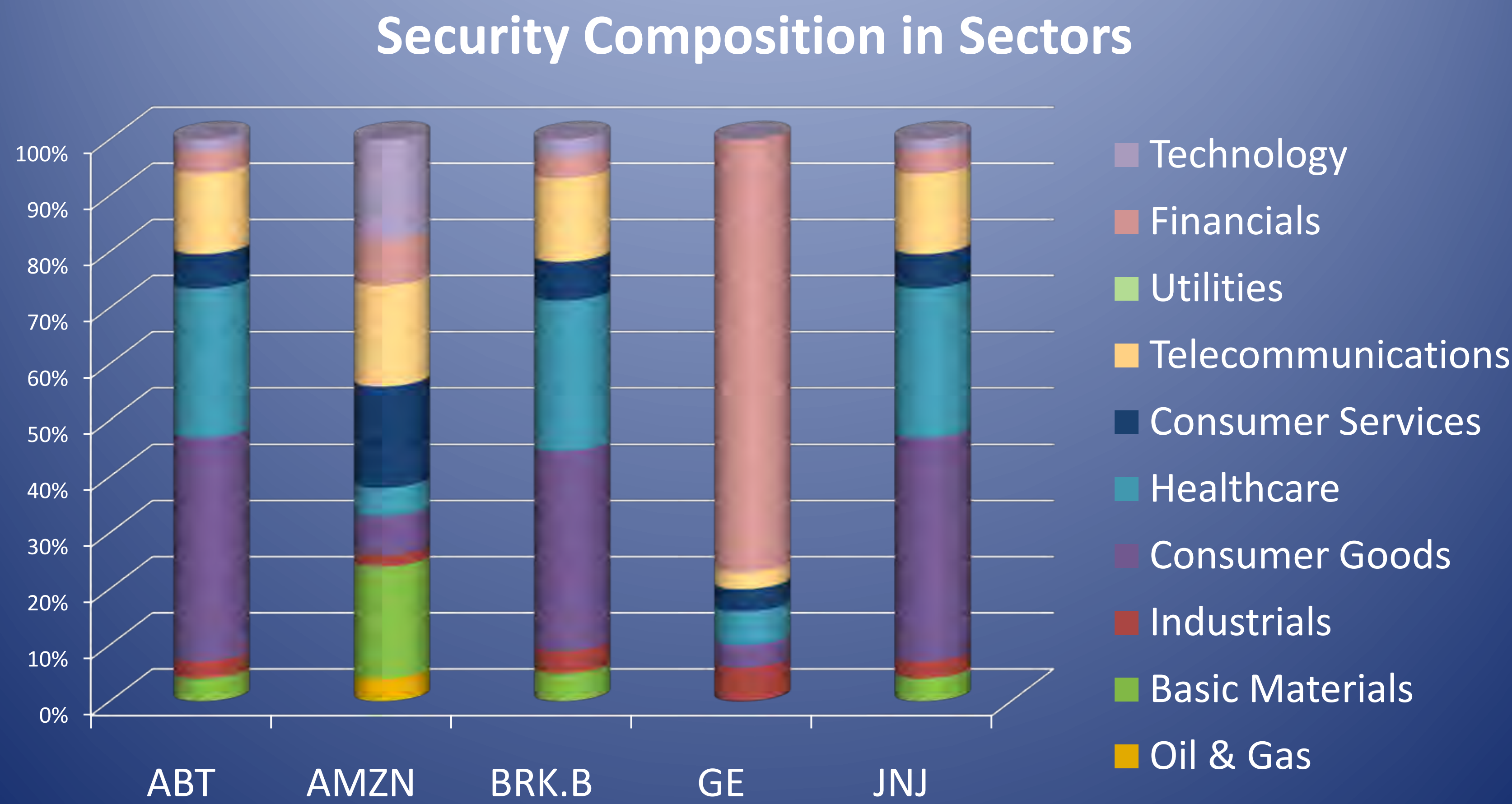
Gaussian Mixture Model



Experiment 5:

5 Years Returns, Fundamentals & β

Gaussian Mixture Model



Northfield
INFORMATION SERVICES, INC.

Closing Remarks

- It works with different distributions
  - Here, deviations from cluster centers were Gaussian
  - Can easily do the same assuming deviations $\sim e^{-\lambda|x|}$, the distribution associated with median. (Gaussian is mean)

- Clustering helps identify what a security is, i.e. what alpha model to use for it

- Switching from applying filters to thinking about underlying mathematical models
  - gives you your own custom tool set
  - makes understanding what something does infinitely easier

Northfield
INFORMATION SERVICES, INC.