Clustering Mixed-Type Data

Speaker: Alex Foss

Research Group: Modern Statistical Learning and Inference Methods,
Prof. M. Markatou

June 1, 2016
UB Computational Sciences Club

Areas of Scientific Inquiry

Major Field: Biostatistics

- Clustering mixed-type data
- Mixture models
- High-dimensional data analysis

Cognate Field: Bioinformatics

- High throughput data
- Imaging data
- Data synthesis; integrative analysis

Statistical Science: Interdisciplinary Approach

(**Remark**: My own path to statistical research)

Interface with biological sciences

- Real-world scientific questions are the best inspiration for meaningful statistical research
- Our central goal is to develop tools for generating high-quality scientific evidence
- **This talk**: Inspired by the problem of grouping subjects using blended 'omics data sets

Interface with computational sciences

- A statistical technique must have a high-quality software implementation to be useful
- Analysis of algorithms
- Collection and analysis of large data sets
- Machine learning, artificial intelligence, data mining
- **This talk**: Complexity analysis, R package development, Map-Reduce computing model

Outline

Problem Statement

We seek to cluster patients into coherent subgroups using mixed-type data.

- **Mixed-type data** refers to a combination of continuous variables (e.g. blood pressure, age) and categorical variables (e.g. race, diagnostic category)
- **Cluster analysis** is a set of methods that identify groups ("clusters") of similar units in a data set

Context

Why does this problem matter?

- Precision medicine: an approach to clinical medicine involving specialized treatments for subgroups of patients
- These subgroups are generally identified through data-driven techniques
- The data are drawn from large and often heterogeneous combinations of sources
  - Demographic information: age, sex, race/ethnicity
  - Diagnostic tests: tumor size, metastasized Y/N
  - High-throughput "omics" data sets: continuous mRNA expression, categorical SNP data

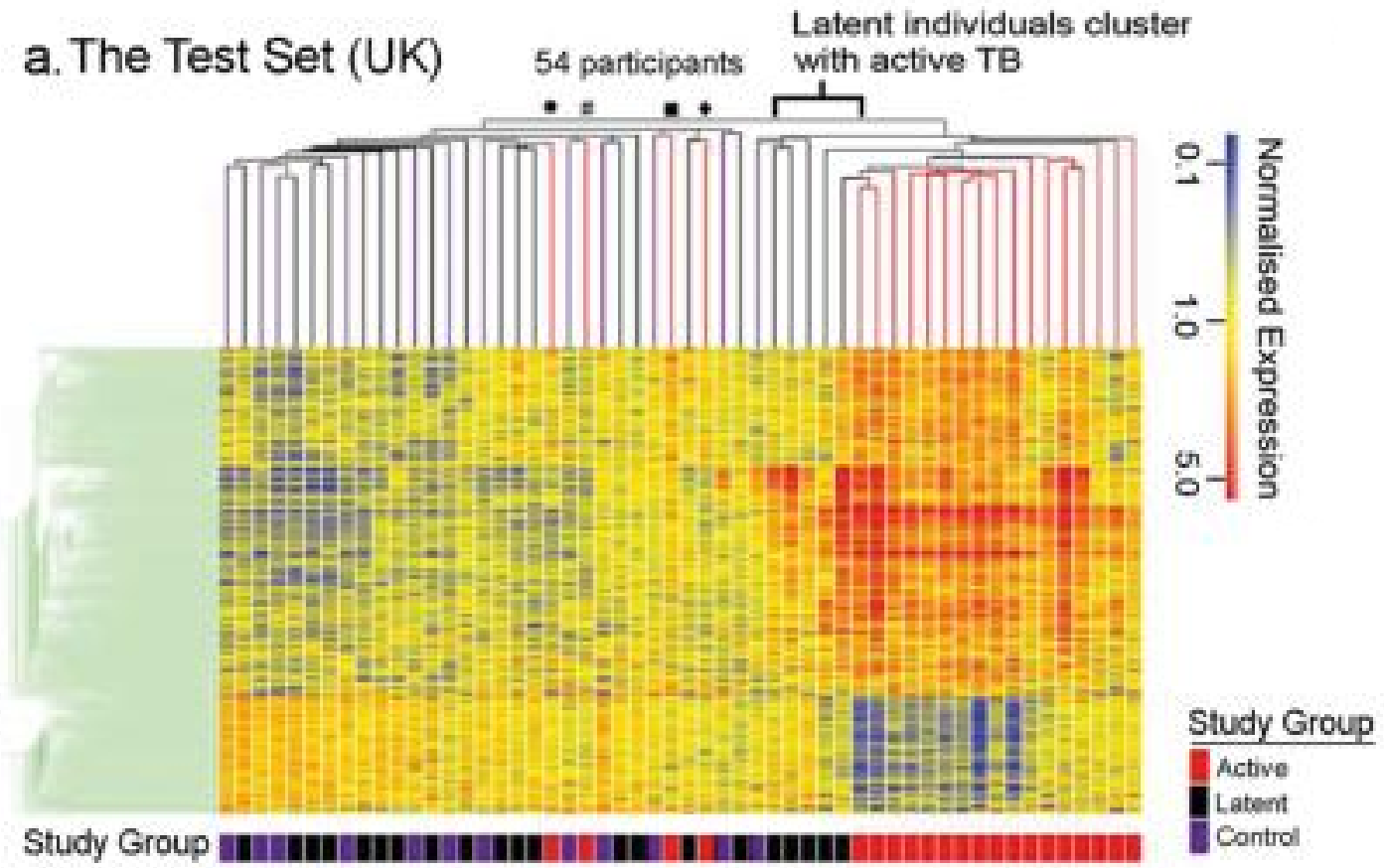


Figure from Berry et al., *Nature* (2010) [1]

Scientific question

Data set: Prostate cancer patients from Byar & Green[1]

- Demographic info
- Clinical measurements (blood pressure, family history)
- Blood chemistry measurements (e.g. serum Hb)
- Tumor morphology, coded EKG results
- Survival

Goals for cluster analysis of this data set:

- Data summary – identify similarities in variables/patients that shed light on a complex data set
- Hypothesis generation – patterns/trends for future study
- Patient matching – groups of homogeneous patients to serve as test/control subjects
- Predict clinical outcomes

[1] Byar & Green, *Bulletin du Cancer* (1980) [2]

Challenges of Identifying Subgroups with Mixed Data

As available data sets are larger and more abundant, mixed-type data sets are increasingly unavoidable. Associated challenges include:

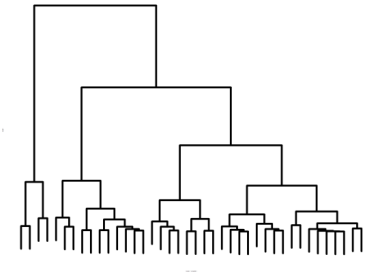
- Existing strategies for clustering mixed data have significant limitations/weaknesses
  - Discretizing continuous variables
  - Dummy coding categorical variables
  - Reliance on user-specified weights
  - Reliance on unrealistic distributional assumptions
- Most existing approaches do not balance the contribution of continuous and categorical data
  - Poor usage of information & less useful clusters
  - Solution tends to be dominated by one variable type

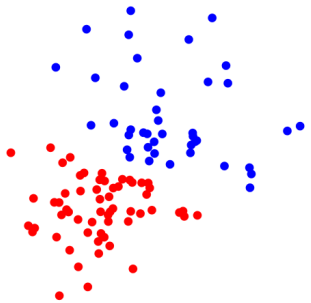
Outline

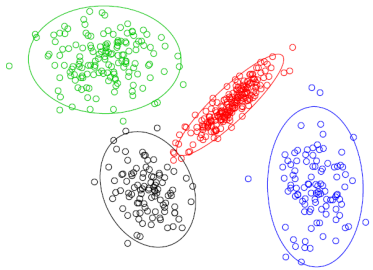
A Taxonomy of Clustering Methods

A Taxonomy of Clustering Algorithms

- **Hierarchical methods**
  - Agglomerative (e.g. AGNES, linkage methods)
  - Divisive (e.g. DIANA)
- **Partitioning methods**
  - Hard partitioning (e.g. standard k-means, PAM)
  - Soft partitioning (e.g. soft k-means)
- **Model-based**
  - Finite mixture models







Four Methods for Clustering Mixed Data

- Discretization
- Numerical Coding
- Hybrid Distance
- Parametric

Discretization Method

Procedure:

1. Discretize all continuous variables
2. Replace continuous variables with discretized version
3. Use a clustering technique suitable for categorical data

Problem: Clustering performance can suffer greatly if discretization is not optimal.
No general method for identifying optimal cut-points

Number of Cut-Points	Mean ARI	Monte Carlo Standard Error
1	0.78	0.003
2	0.27	0.003
3	0.23	0.002
4	0.21	0.002

Numerical Coding Methods

Procedure:

1. Convert categorical variables into continuous (e.g. dummy coding)
2. Replace all categorical variables with coded version
3. Use a clustering technique suitable for continuous data

Problem: Any fixed coding strategy will be sub-optimal in the general case

Numerical Coding Methods

Consider sampling from (V, W), where

$$V = \begin{cases} Y_1 \text{ with probability } \pi \in [0,1] \\ Y_2 \text{ with probability } 1-\pi \end{cases}$$

where Y_1, Y_2 follow continuous distributions with means μ and 0, and variances σ_1^2 and σ_2^2; V is z-normalized before clustering.
Let W represent a 0-1 dummy coded mixture:

$$W = \begin{cases} B_1 \text{ with probability } \pi \\ B_2 \text{ with probability } 1-\pi \end{cases}$$

with $B_1 \sim Bern(p_1)$ and $B_2 \sim Bern(p_2)$.

Consider the squared Euclidean distance between population 1 and 2:

- The continuous contribution has expectation > 1 for $\sigma_1 \neq \sigma_2$
- The continuous contribution has expectation > 2 for $\sigma_1 = \sigma_2$
- The categorical contribution has expectation < 1

$\Rightarrow$ **Unbalanced treatment of continuous and categorical.**

Perhaps categorical variable should be upweighted, i.e. 0–2 dummy coding instead of 0–1?
This is also ineffective in the general case.

Consider sampling from random vector (V, W):

$$\begin{cases} V \sim N(0,1), & W \sim Multin(n=1, \mathbf{p}=(0.5-\epsilon, 0.5-\epsilon, \epsilon, \epsilon)), & \text{if drawn from pop. 1} \\ V \sim N(\mu,1), & W \sim Multin(n=1, \mathbf{p}=(\epsilon, \epsilon, 0.5-\epsilon, 0.5-\epsilon)), & \text{if drawn from pop. 2} \end{cases}$$

with $\epsilon \in [0, 0.25)$.

		K-means 0–1 coding			K-means 0–2 coding			FMM		
		Categorical Overlap			Categorical Overlap			Categorical Overlap		
		1%	15%	30%	1%	15%	30%	1%	15%	30%
Continuous Overlap	1%	0.99	0.99	0.98	0.99	0.89	0.80	1.00	0.99	0.99
	15%	0.85	0.81	0.78	0.96	0.81	0.66	0.98	0.87	0.81
	30%	0.64	0.60	0.56	0.89	0.71	0.48	0.97	0.78	0.66

$\Rightarrow$ Any fixed coding strategy will be sub-optimal in the general case.

Hybrid Distance

Procedure:

- Use a distance suitable for mixed data (e.g. Gower)
- Use a clustering method that only depends on the pairwise distances between data points (e.g. DIANA, AGNES, PAM)

Problem (same as numerical coding): Requires choice of arbitrary weighting factor controlling continuous/categorical contribution

Example: The k-prototypes distance function[2] defines the distance between observations $(\mathbf{v}_1^T, \mathbf{w}_1^T)^T$ and $(\mathbf{v}_2^T, \mathbf{w}_2^T)^T$ is

$$d_{con}(\mathbf{v}_1, \mathbf{v}_2) + \gamma d_{cat}(\mathbf{w}_1, \mathbf{w}_2)$$

[2] Proposed by Z Huang, *Data Mining and Knowledge Discovery* (1998) [6]

Hybrid Distance Method: Modha-Spangler Weighting

Modha and Spangler (2003) [8] propose a way to estimate the optimal weighting factor γ for continuous versus categorical contribution

Procedure:

1. **Within cluster distortion**: the sum of distances from the cluster centroid to each point in the cluster
2. **Total distortion**: sum of all distances to the overall centroid across the entire data set
3. **Between cluster distortion**: total minus within cluster distortion
4. Calculate within-to-between distortion ratio separately for continuous and categorical variables; call them Q_{con} and Q_{cat}
5. Brute force search for the γ that minimizes $Q_{con} \times Q_{cat}$

Hybrid Distance Method: Modha-Spangler Weighting

- Pro: Data-drive weight selection
- Pro: Better than arbitrary weighting, but still often overemphasize continuous
- Con: Cannot up- or down-weight individual variables
- Con: Unstable performance; may minimize $Q_{con} \times Q_{cat}$ by allocating each categorical level to its own cluster

Hybrid Distance Method: Modha-Spangler Weighting

Simulation conditions:

- Two continuous variables (mixture of normals)
  Overlap $\in \{1\%, 15\%, 30\%, 45\%\}$
- Five binary variables
  - One with nearly perfect separation by cluster
  - Four with no relationship to clusters
- Two clusters
- $N = 500$
- 500 Monte Carlo samples per condition
- Methods: k-means with dummy coding, finite mixture model (FMM), Modha-Spangler

Hybrid Distance Method: Modha-Spangler Weighting

Continuous Overlap	FMM ARI	k-means ARI	Modha-Spangler ARI
0.01	1.00	1.00	1.00
0.15	0.99	0.97	0.99
0.30	0.98	0.85	0.94
0.45	0.98	0.67	0.84

Statistical Clustering Methods

Example: A Gaussian mixture model[3]: the data vectors $\mathbf{X}_i$ to be clustered are assumed to be drawn from one of G possible Gaussian distributions:

$$\mathbf{X}_i \sim \begin{cases} N(\boldsymbol{\mu}_1, \Sigma_1), & \text{with probability } \pi_1 \\ N(\boldsymbol{\mu}_2, \Sigma_2), & \text{with probability } \pi_2 \\ \quad\vdots & \\ N(\boldsymbol{\mu}_G, \Sigma_G), & \text{with probability } \pi_G \end{cases}$$

where $\sum_{g=1}^{G} \pi_g = 1$ and $0 < \pi_g < 1 \; \forall \; g$. The density of $\mathbf{X}_i$ is $f_{\mathbf{X}_i}(t) = \sum_{g=1}^{G} \pi_g f(t;\, \boldsymbol{\mu}_g, \Sigma_g)$, where f is the normal density.

[3] Hastie et al., *The Elements of Statistical Learning* [4]

Statistical Clustering Methods

- For mixed data, use e.g. a normal-multinomial mixture model
- Pro: It can successfully balance continuous and categorical contribution
- Pro: It can identify and avoid uninformative variables.
- Pro: No weight specification necessary
- Con: Affected by violations to parametric assumption

Clustering Mixed Data

Summary:

- How to equitably balance continuous and categorical contribution is a central problem (numerical coding, hybrid distance methods)
- Methods for equitable treatment tend to create vulnerability to uninformative variables (Modha-Spangler, ensemble)
- Statistical methods solve both problems under certain conditions
- Statistical methods are susceptible to violations of parametric assumptions
- **Our novel clustering method, KAMILA, addresses all of these problems**

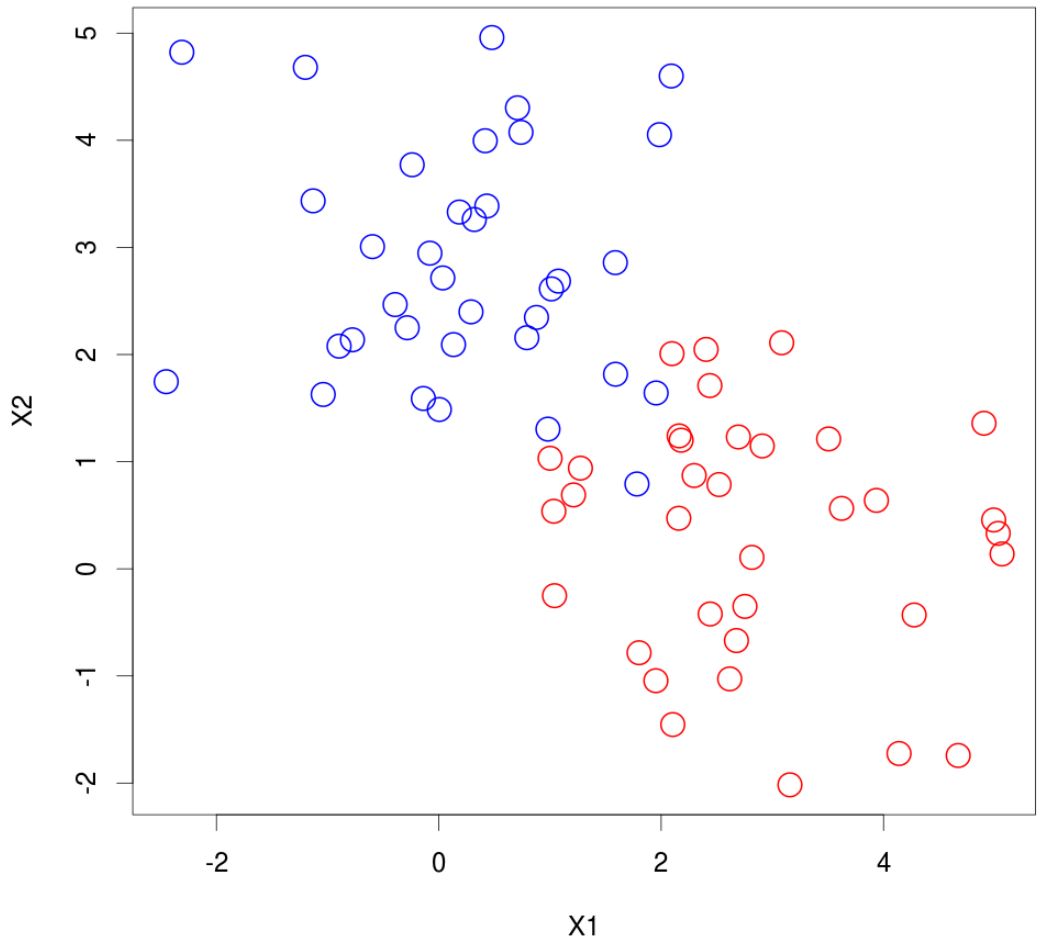
Outline

KAMILA: KAy-Means for MIxed LArge datasets

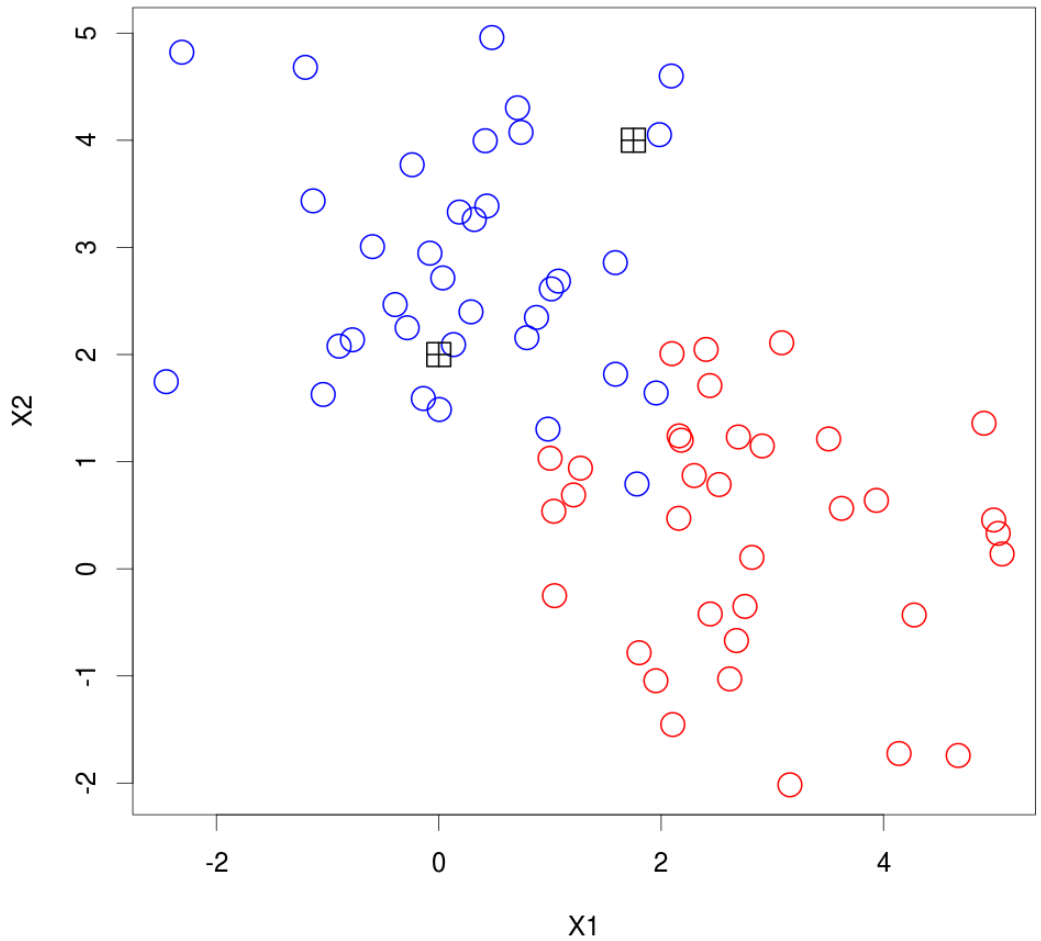
We seek to unify two lines of research from the separate fields of statistics and CS:

- Optimal properties of model-based clustering methods
- Relaxed distributional assumptions based on spherical kernel density estimation
- Can accommodate any spherical or elliptical cluster distribution

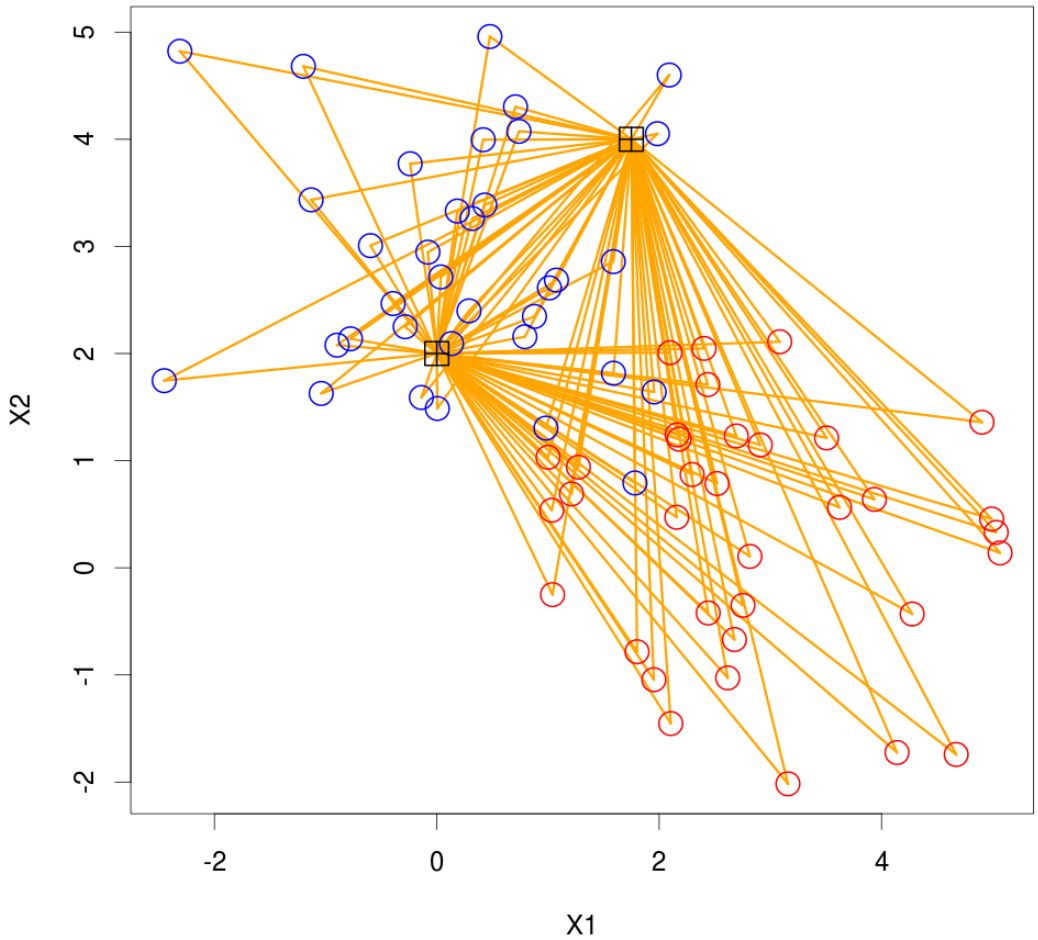
KAMILA Algorithm



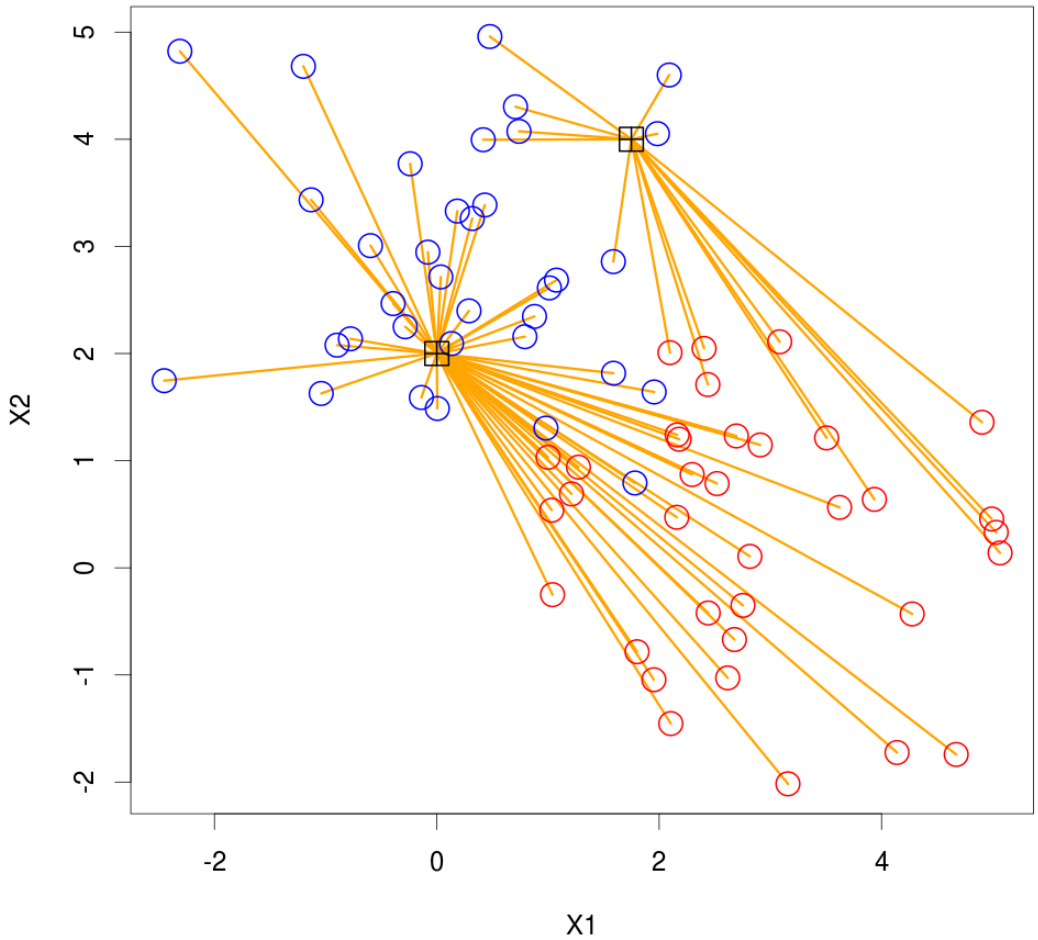
KAMILA Algorithm



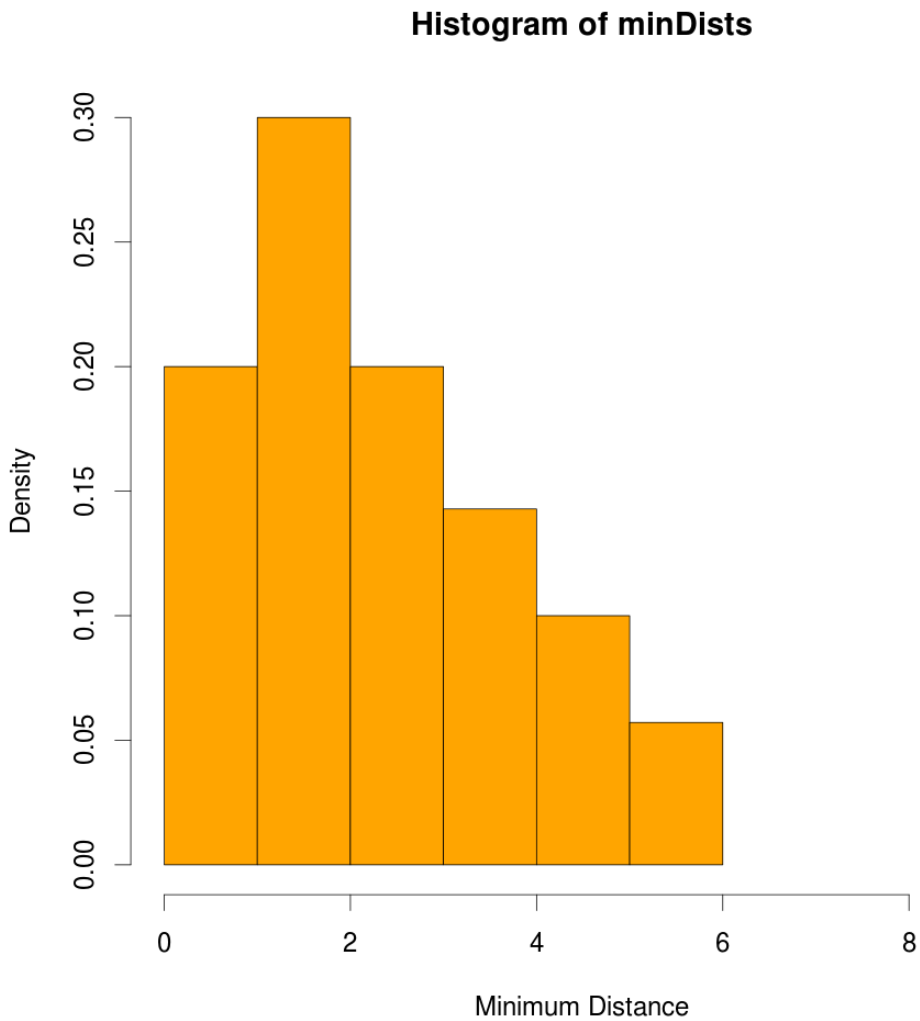
KAMILA Algorithm



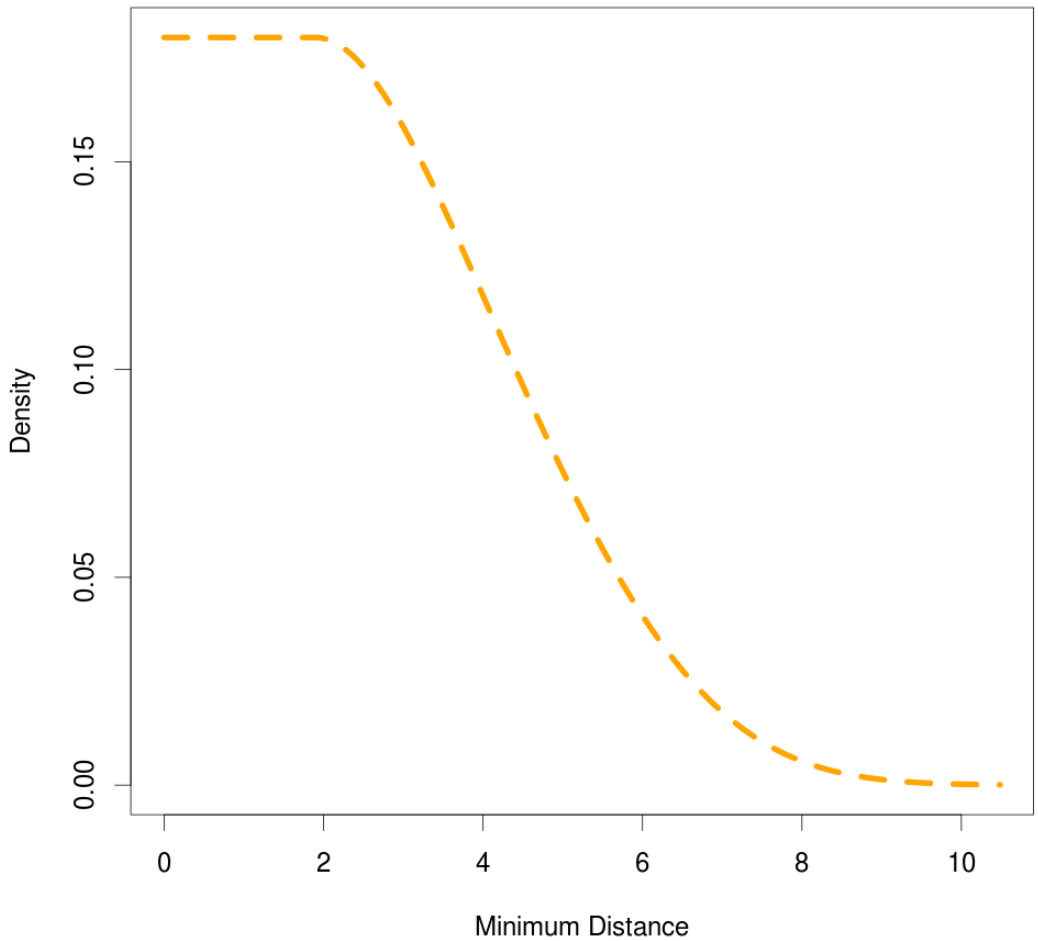
KAMILA Algorithm



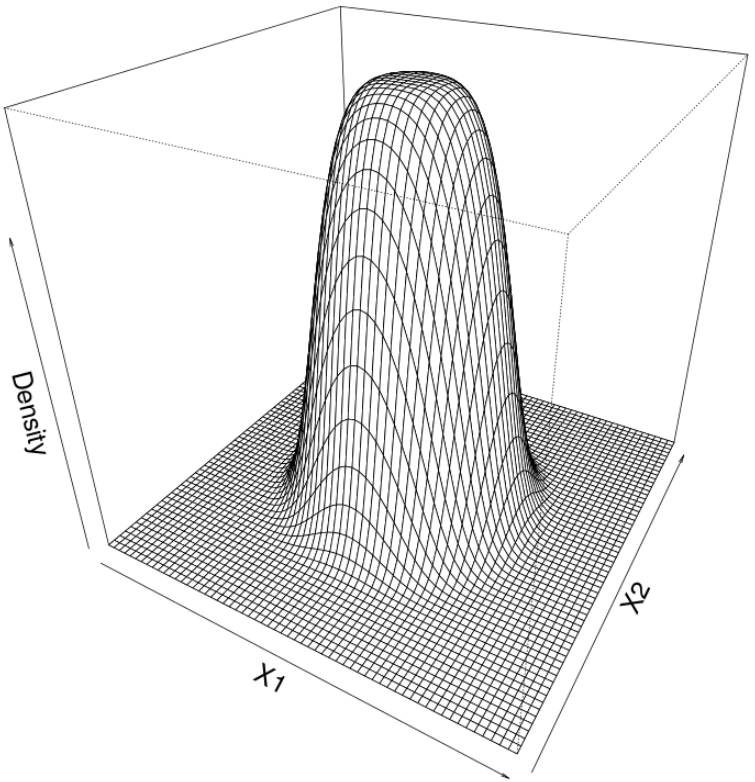
KAMILA Algorithm



KAMILA Algorithm



KAMILA Algorithm



Statistical Innovations: Radial KDE

Continuous variables are modeled as a mixture of elliptical distributions which are determined nonparametrically from the data.

Proposition 2 in Foss et al.[4] states that if $\mathbf{X} \in \mathbb{R}^p$ follows a spherically symmetric distribution with center μ, then

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{f_R(r)\,\Gamma(\frac{p}{2}+1)}{p r^{p-1} \pi^{p/2}},$$

where $r = \sqrt{(\mathbf{x}-\mu)^T(\mathbf{x}-\mu)}$, $R = \sqrt{(\mathbf{X}-\mu)^T(\mathbf{X}-\mu)}$, and f_R is the probability density of R.
Note: using a scaling matrix Σ this result can be easily extended to elliptical distributions.

[4] Foss, Markatou, et al., *Machine Learning* (Revision Submitted) [3]

Statistical Innovations: Radial KDE

- Construct $\hat{f}_R$ using a (univariate) kernel density estimation scheme, and substitute it for f_R above to obtain $\hat{f}_X$

$$\hat{f}_R^{(t)}(r) = \frac{1}{Nh^{(t)}} \sum_{\ell=1}^{N} k\left(\frac{r - r_\ell^{(t)}}{h^{(t)}}\right),$$

  where $k(\cdot)$ is a kernel function and $h^{(t)}$ the corresponding bandwidth parameter at iteration t. The Gaussian kernel is currently used, with bandwidth $h = 0.9An^{-1/5}$, where $A = \min(\hat{\sigma}, \hat{q}/1.34)$, $\hat{\sigma}$ is the sample standard deviation, and $\hat{q}$ is the sample interquartile range[5].
- This yields an approximation of f_X that is restricted to have spherical contour lines.
- It avoids the drawbacks of a multivariate kernel density estimator (computationally expensive and potential to overfit data points).

[5] Silverman, *Density Estimation* (1986) [9, p. 48, equation 3.31]

Statistical Innovations: Data Model

Data are IID draws from $(\mathbf{V}^T, \mathbf{W}^T)^T_{(P+Q)\times 1}$

- $\mathbf{V} \sim \sum_{g=1}^{G} \pi_g h(\mathbf{v}; \mu_g, \Sigma_g)$, where π_g denotes the prior probability of observing the g^{th} cluster, μ_g denotes its centroid, and Σ_g its scaling matrix.
- Density h is elliptically symmetric as described above
- $\mathbf{W} \sim \sum_{g=1}^{G} \pi_g \prod_{q=1}^{Q} multin(w_q; \theta_{gq})$, where θ_{gq} is the multinomial parameter vector for the g^{th} component of the q^{th} categorical variable.
- Conditional independence between $\mathbf{V}$ and $\mathbf{W}$ within the g^{th} cluster $\forall\, g$
- In the t^{th} iteration, point i is assigned to the cluster g that maximizes

$$\log\left[\hat{f}^{(t)}_{\mathbf{V}}(d^{(t)}_{ig})\right] + \log\left[\prod_{q=1}^{Q} multin(w_q; \hat{\theta}^{(t)}_{gq})\right]$$

Statistical Innovations: Algorithm

Algorithm 1 KAMILA Clustering

for User-specified number of initializations **do**

Initialize $\hat{\mu}_g^{(0)}, \hat{\theta}_{gq}^{(0)}\ \forall g, q$

repeat

PARTITION STEP

$d_{ig}^{(t)} \leftarrow \text{dist}(\mathbf{v}_i, \hat{\mu}_g^{(t)})$

$r_i^{(t)} \leftarrow \min_g(d_{ig}^{(t)})$

$\hat{f}_{\mathbf{V}}^{(t)} \leftarrow \text{RadialKDE}(\mathbf{r}^{(t)})$

$c_{ig}^{(t)} \leftarrow \widehat{Pr}(\mathbf{w}_i \,|\, \text{observation } i \in \text{population } g)$

$H_i^{(t)}(g) \leftarrow \log\left[\hat{f}_{\mathbf{V}}^{(t)}(d_{ig}^{(t)})\right] + \log\left[c_{ig}^{(t)}\right]$

Assign observation i to population $\text{argmax}_g\{H_i^{(t)}(g)\}$

ESTIMATION STEP

Calculate $\hat{\mu}_g^{(t+1)}$ and $\hat{\theta}_{gq}^{(t+1)}$

until Convergence

$ObjectiveFun \leftarrow \sum_{i=1}^{N} \max_g\{H_i^{(final)}(g)\}$

end for

Output partition that maximizes $ObjectiveFun$

Statistical Innovations: Algorithm

- Initialization: sample G points randomly from the data set for continuous coordinates; randomly generate $\hat{\theta}_{gq}^{(0)}$ from the flat Dirichlet distribution
- Selecting # of clusters
  - We use the prediction strength method of Tibshirani & Walther[6]
  - Based on measuring consistency of clusters in multiple cross-validation runs
- Stopping rule: terminate run when group membership is unchanged in two successive iterations
- Multiple initializations run to completion to avoid local maxima

[6] Tibshirani & Walther, *J Comp. and Graphical Stat.* (2005) [10]

Statistical Innovations: Identifiability

Holzmann et al.[7] discuss identifiability of mixtures of elliptical distributions of the form

$$f_{\alpha,p}(\mathbf{x}) = |\Sigma|^{-1/2} f_p\left[(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}); \boldsymbol{\theta}\right]$$

where $\mathbf{x} \in \mathbb{R}^p$, $\boldsymbol{\alpha} = (\boldsymbol{\theta}, \boldsymbol{\mu}, \Sigma) \in \mathcal{A}^p \subset \mathbb{R}^{k\times p\times p(p+1)/2}$, $f_p : [0,\infty) \to [0,\infty)$ is a density generator, that is, a nonnegative function such that $\int f(\mathbf{x}^T\mathbf{x}; \boldsymbol{\theta})\mathrm{d}\mathbf{x} = 1$.

Theorem 2, Holzmann et al.

Let $f_p(\cdot;\theta)$, $\theta \in \Theta$, be a parametric family of density generators for spherically symmetric distributions in $\mathbb{R}^p$. Let $\mathcal{C} = \Theta \times (0,\infty) \times \mathbb{R}$ and let $\gamma_j = (\theta_j, a_j, b_j) \in \mathcal{C}$ for $j = 1, 2$. Suppose there exists a total ordering $\preceq$ on the set $\mathcal{C}$ such that $\gamma_1 \prec \gamma_2$ implies

$$\lim_{u\to\infty} \frac{f_p(a_2u^2 + b_2u + c_2;\, \theta_2)}{f_p(a_1u^2 + b_1u + c_1;\, \theta_1)} = 0 \quad \textit{for} \quad c_1, c_2 \in \mathbb{R}.$$

Then finite mixtures from the family $\{f_{\alpha,p} : \alpha = (\theta, \mu, \Sigma) \in \mathcal{A}^p\}$ of elliptical distributions in $\mathbb{R}^p$ are identifiable.

[7] Holzmann et al., *Scandinavian J. of Statistics* (2006) [5]

Statistical Innovations: Identifiability

Proposition 3, Foss, Markatou et al. [3]

A1. The kernel $k(\cdot)$ is a positive function such that $\int k(u)du = 1$, $\int u\,k(u)du = 0$, and $\int u^2\,k(u)du > 0$.

A2. The kernel function $k(\cdot)$ is a continuous, monotone decreasing function such that $\lim_{u\to\infty} k(u) = 0$.

A3. The kernel $k(\cdot)$ is such that $\lim_{z\to\infty} \frac{k(z\,\gamma_2)}{k(z\,\gamma_1)} = 0$, where γ_1, γ_2 are constants such that $\gamma_2 > \gamma_1$.

A4. The number of clusters G is fixed and known, with different centers $\boldsymbol{\mu}_j$, $j = 1, 2, ..., G$.

A5. The density functions of the different clusters come from the same family of distributions and differ in terms of their location parameters, i.e. they are $f(\mathbf{v} - \boldsymbol{\mu}_j)$, $j = 1, 2, ..., G$.

Under assumptions A1–A5 the KAMILA density generator satisfies the conditions of Theorem 2 of Holzmann et al. [5].

Time Complexity

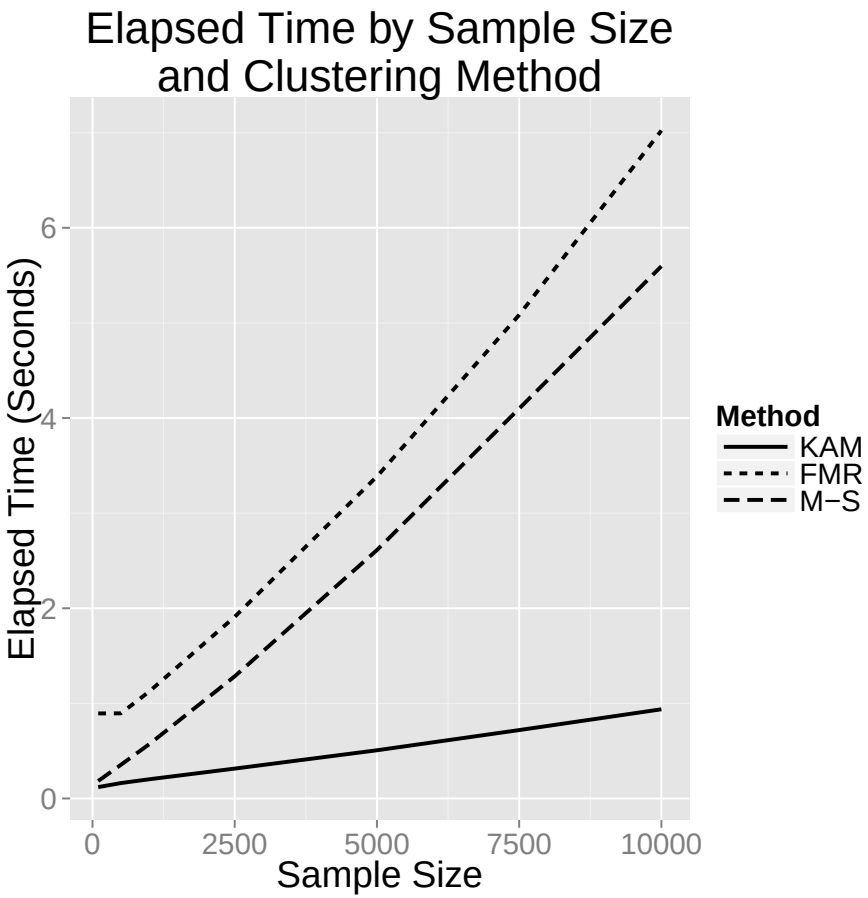
Time complexity is linear in n with an upper bound on the number of iterations:

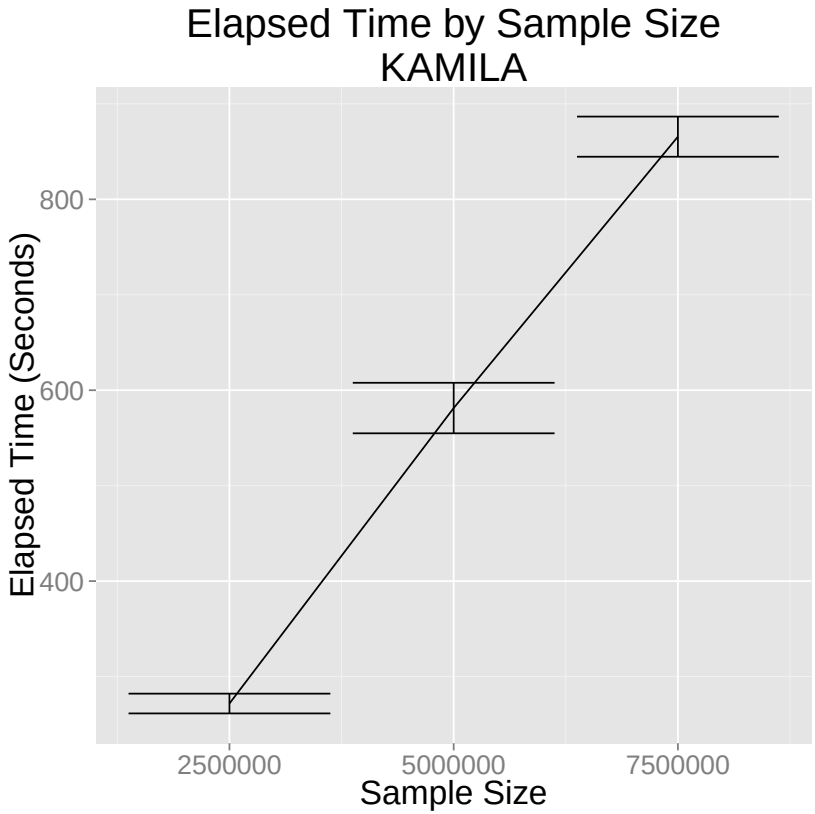
$$\mathcal{O}(knit)$$

- k = number of clusters
- n = sample size
- i = number of initializations
- t = upper bound on iterations

A realistic theoretical bound with t unrestricted is difficult even in the simpler case of k-means. (See Arthur et al., (2011), *J. ACM* **58**(5)).

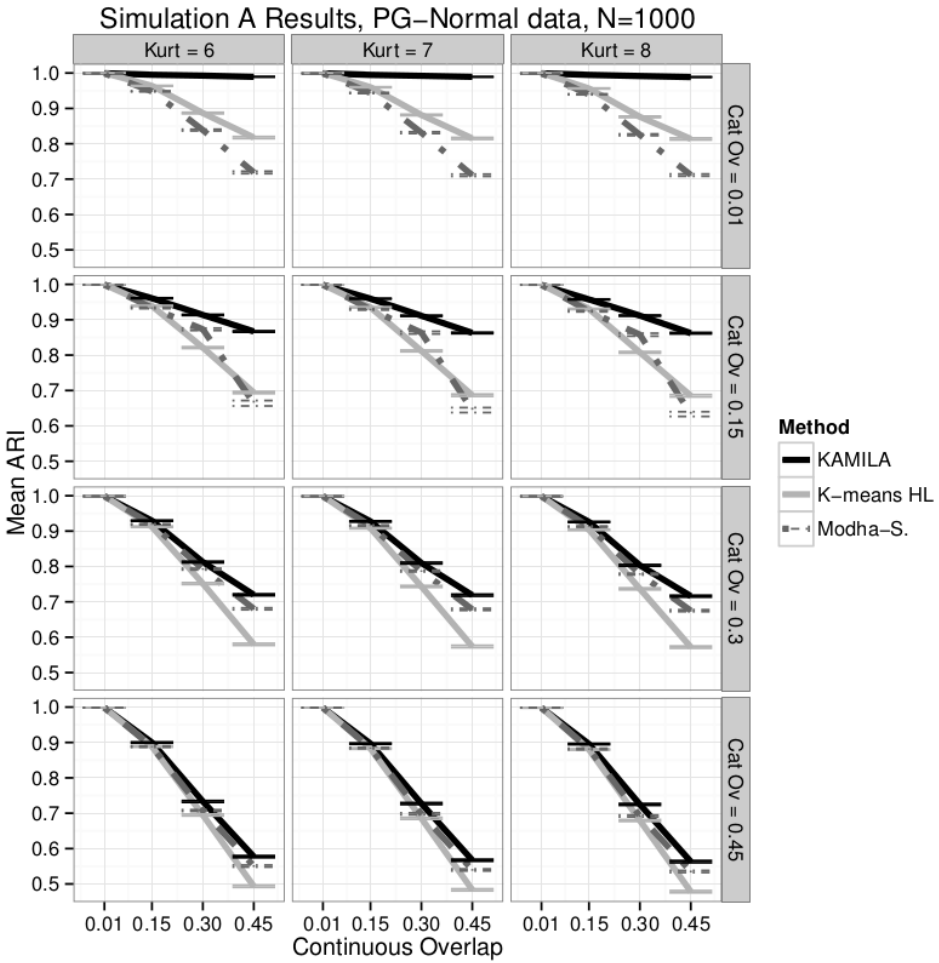
Time Complexity





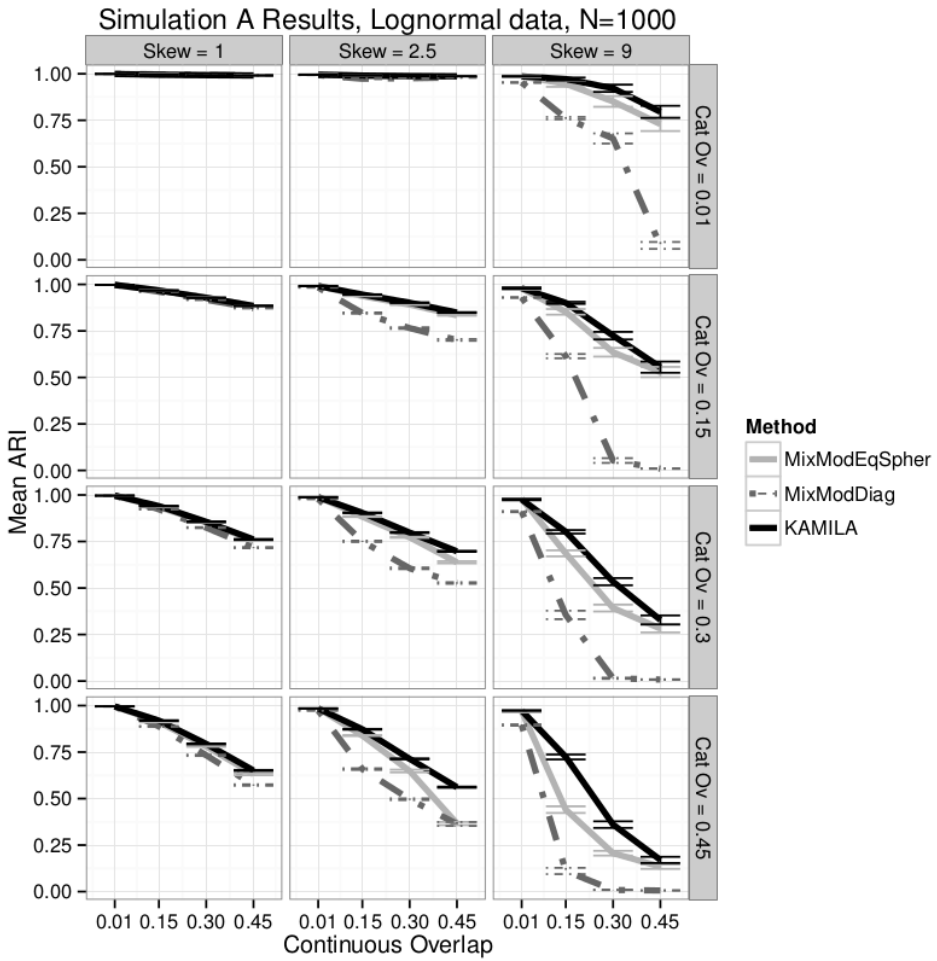
KAMILA Performance

Fig. 4 Simulation A, p-generalized normal data. Direct comparison of ARI between Modha-Spangler weighting, Hennig-Liao weighting, and KAMILA by kurtosis of the continuous variable, and by continuous and categorical overlap levels. Results shown for $N = 1000$.



KAMILA Performance

Fig. 5 Simulation A, lognormal data. Direct comparison of ARI between the finite mixture models and KAMILA by skewness of the continuous variable, and by continuous and categorical overlap levels. Results shown for $N = 1000$.



Software Development

We have developed an R package containing a suite of tools for clustering mixed-type data

- Clustering techniques: KAMILA, Modha-Spangler
- Weighting techniques: Hennig-Liao, MEDEA, Burnaby

We have developed a Hadoop implementation of KAMILA

- Designed for clustering very large data sets stored on distributed file systems
- Map-Reduce programming model

Outline

Prostate Cancer Dataset

Dataset:

- Data sampled from a population of prostate cancer patients (Byar & Green, 1980 [2])
- Continuous vars: Blood pressure, serum hemoglobin, size of tumor, tumor severity, serum prostatic acid phosphatase
- Categorical vars: Activity level, family history of cardiovascular disease (Y/N), coded EKG results, bone metastases (Y/N), disease stage (3 or 4)

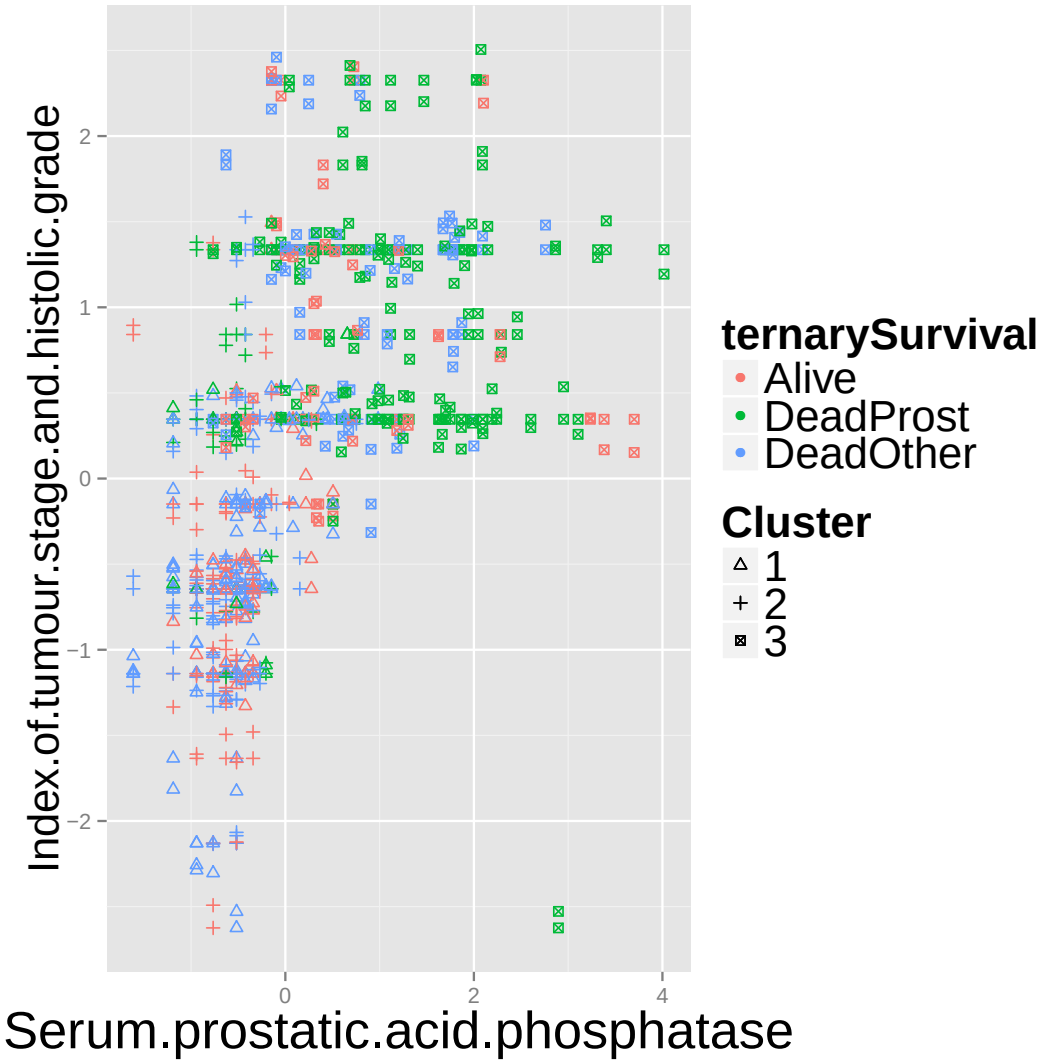
Objectives:

1. Identify meaningful subgroups of patients
2. Identify variables relevant to outcome of interest (survival)
3. Generate new hypotheses

Analysis

```r
# Load data
data(Byar, package='clustMD')

# Log transform SPAP
SPAP <- 'Serum.prostatic.acid.phosphatase'
Byar[,SPAP] <- log(Byar[,SPAP])

# Extract variables of interest; NOT SURVIVAL
conVars <- Byar[,c(5,6,8:11)]
catVars <- data.frame(lapply(Byar[,c(3,4,7,12,13)],factor))

# z-normalize continuous variables
conVars <- as.data.frame(scale(conVars))

# KAMILA clustering
set.seed(1234)
results <- kamila(conVars, catVars, numClust=3,
        numInit=10, maxIter=20)
```

Prostate Clustering Methodology and Results



Implications of Prostate Cancer Analysis

- Without any variable selection or modeling work, we have identified a potential classification strategy for prostate cancer mortality ($\chi^2(4) = 99.7, p = 1.12 \times 10^{-20}$)
- Patients who died due to unrelated causes appear to be more similar to those alive VS patients dead due to prostate cancer

	Alive	DeadProst	DeadOther
Cluster 1	27%	9%	64%
Cluster 2	38%	12%	50%
Cluster 3	20%	52%	28%

Table : Cluster number vs outcome, row percentages

	No	Yes
Cluster 1	116	2
Cluster 2	188	0
Cluster 3	94	75

Table : Cluster vs bone metastasis, raw counts

	Cancer Stage	
	3	4
Cluster 1	81%	19%
Cluster 2	93%	7%
Cluster 3	2%	98%

Table : Cluster vs cancer stage, row percentages

	History of CVD	
	No	Yes
Cluster 1	36%	64%
Cluster 2	62%	38%
Cluster 3	64%	36%

Table : Cluster vs history of cardiovascular disease, row percentages

Outline

Impact of Current Work

- Cluster analysis is a key tool in precision medicine.
- Existing tools for clustering mixed-type data are deficient.
  - Discretization overly dependent on cut-point selection
  - Numerical coding, hybrid distance do not balance continuous and categorical effectively
  - Statistical methods solve these problems, but require the strongest assumptions which may be violated
- Our proposed method doesn't suffer from the limitations of existing methods
  - Sensible balancing of continuous vs categorical contribution means that information is not senselessly discarded
  - Lack of weight specification facilitates reproducibility and ease of use
  - Computational efficiency allows KAMILA to be used with very large data sets

Technical challenges

- Convergence of the algorithm
  - Similar in structure to k-means, EM algorithm
  - Each step of the estimation algorithm increases the objective function, but can get trapped in local maxima
- Optimal kernel selection
  - Bias/variance of kernel density estimator, asymptotics
  - Derive optimal kernel/bandwidth for clustering
- Initialization
  - We have found that initializing centroids using random data points outperforms the use of random points generated uniformly from the parameter space
- Stopping rule
  - Currently running until group membership remains unchanged
  - Consider numerical techniques to stop earlier
- Extensions to longitudinal data (time-dependencies)

Relevant papers

- Foss, Markatou, Ray & Heching. A Semiparametric Method for Clustering Mixed Data. *Machine Learning* (Revision submitted).

In progress

- Foss & Markatou. Clustering mixed-type data in R and Hadoop. *Journal of Statistical Software* (In preparation).
- Foss, Markatou, & Ray. Clustering mixed-type data: Review and synthesis. *International Statistical Review* (In preparation).
- Foss & Markatou. Convergence of the KAMILA clustering algorithm. *Neural Information Processing* (In preparation).

Future Directions

- Extend semiparametric clustering techniques to alternate clustering models
  - Linkage-based clustering
  - mean-shift algorithms (density-based clustering)
- Extend to a mixture of regressions framework
  - Clusters the data while simultaneously estimating within-cluster regression equations
  - Regression model fitting informs clustering stage
- Performance in high-dimensional data sets
  - The radial kernel density estimation step effectively reduces the dimension of the continuous variables
  - KAMILA performs well when the number of continuous variables P is greater than n
  - Investigate asymptotic properties when $P \to \infty$ and n is fixed

Q & A

[1] M Berry, C Graham, and F McNab. An interferon-inducible neutrophil-driven blood trascriptional signature in human tuberculosis. *Nature*, 466:973–977, 2010.

[2] D Byar and S Green. The choice of treatment for cancer patients based on covariate information: applications to prostate cancer. *Bulletin du Cancer*, 67: 477–490, 1980.

[3] A Foss, M Markatou, B Ray, and A Heching. A semiparametric method for clustering mixed data. *Machine Learning*, X(X):X–X, Revision Submitted.

[4] T Hastie, R Tibshirani, and J Friedman. *The Elements of Statistical Learning, Second Ed.* Springer, New York, USA, 2009.

[5] H Holzmann, A Munk, and T Gneiting. Identifiability of finite mixtures of elliptical distributions. *Scandinavian Journal of Statistics*, 33(4):753–763, 2006.

[6] Z Huang. Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining and Knowledge Discovery*, 2(3):283–304, 1998.

[7] L Hubert and P Arabie. Comparing partitions. *Journal of Classification*, 2(1): 193–218, 1985.

[8] DS Modha and WS Spangler. Feature weighting in k-means clustering. *Machine Learning*, 52(3):217–237, 2003.

[9] BW Silverman. *Density Estimation*. Chapman and Hall, London, 1986.

[10] R Tibshirani and G Walther. Cluster validation by prediction strength. *J Computational and Graphical Statistics*, 14(3):511–528, 2005.

[11] R Tibshirani, G Walther, and T Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

Background: Distance Measures

There are three types of distances:

- Distances between distributions
- Distances between vectors
- Distances between a distribution and a vector

Examples of distance between two distributions:

- Kullback-Leibler divergence

$$d_{KL}(F_1, F_2) = \int \log\left(\frac{dF_1}{dF_2}\right) dF_1$$

- Mahalanobis

$$d_M(F_1, F_2) = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \Sigma^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)},$$

Background: Distance Measures

Examples of distance between two data vectors:

- **Mahalanobis**

$$d_M(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T \Sigma^{-1} (\mathbf{x}_1 - \mathbf{x}_2)},$$

  where $Var(\mathbf{x}_1) = Var(\mathbf{x}_2) = \Sigma$

- **Euclidean**: Mahalanobis with $\Sigma =$ Identity
- **Matching distance**: For $p \times 1$ categorical vectors $\mathbf{x}$ and $\mathbf{y}$

$$\sum_{i=1}^{p} I\{x_i \neq y_i\}/p$$

  where $I\{\cdot\}$ is the indicator function.

Measuring the Performance of Clustering Algorithms

Two types of measures

- External: Used if true population memberships are **known**
- Internal: Used if true population memberships are **unknown**

Challenges in constructing **external** measures:

- **Accuracy is undefined**
- No clear correspondence between true populations and estimated clusters
- # clusters may not equal # populations

		Cluster		
		A	B	C
True	1	20	0	0
Population	2	0	10	10

Measuring the Performance of Clustering Algorithms

- The Adjusted Rand Index (ARI) is a popular and useful external measure [7]
- Corrected for chance level performance
- ARI $= 1 \Rightarrow$ perfect correspondence
- ARI $= 0 \Rightarrow$ chance level performance

ARI $= 0.23$

	A	B	C
1	10	10	0
2	0	10	10

ARI $= 0.74$

	A	B	C
1	20	0	0
2	0	10	10

ARI $= 0.91$

	A	B	C
1	20	0	0
2	0	18	2

Selecting the Number of Clusters

Choosing the number of clusters G is a difficult problem

- Choose G that optimizes some internal criterion
- Information-based methods (AIC, BIC, QIC etc)
- Gap statistic [11]; based on a simulated "null" distribution
- Prediction strength [10]; based on the concept of generalization error

Two Central Challenges of Clustering Mixed Data

1. How to equitably balance the continuous and categorical contribution to the clustering
2. How to minimize the effects of uninformative variables

Two Challenges of Clustering Mixed Data

Consider two distinct random vectors (V_1, W_1) and (V_2, W_2) defined as

$$\begin{cases} V_1 \sim N(0,1), & W_1 \sim Multin(n=1, \mathbf{p}=(0.45, 0.45, 0.05, 0.05)), \quad \text{with probability } \pi \\ V_1 \sim N(4,1), & W_1 \sim Multin(n=1, \mathbf{p}=(0.05, 0.05, 0.45, 0.45)), \quad \text{with probability } 1-\pi. \end{cases}$$

$$\begin{cases} V_2 \sim N(0,1), & W_2 \sim Multin(n=1, \mathbf{p}=(0.25, 0.25, 0.25, 0.25)), \quad \text{with probability } \pi \\ V_2 \sim N(4,1), & W_2 \sim Multin(n=1, \mathbf{p}=(0.25, 0.25, 0.25, 0.25)), \quad \text{with probability } 1-\pi. \end{cases}$$