Pattern Recognition of Big Nutritional Data in RCT

Jin Wang
University of Massachusetts Dartmouth, Department of Electrical and Computer Engineering
285 Old Westport Road, North Dartmouth, MA, USA 02747
jwang7@umassd.edu

Hua Fang
University of Massachusetts Medical School, Department of Quantitative Health Sciences
368 Plantation Street, Worcester, MA USA01605-0002
hua.fang@umassmed.edu

Honggang Wang
University of Massachusetts Dartmouth, Department of Electrical and Computer Engineering
285 Old Westport Road, North Dartmouth, MA, USA 02747
Hwang1@umassd.

Gin-Fei Olendzki
University of Massachusetts Medical School, Department of Medicine
Health Sciences
55 Lake Avenue North, Worcester, MA 01655, USA
effie.chung@umassmed.edu

Chonggang Wang
Interdigital
781 Third Avenue
King of Prussia, PA 19406, USA
cgwang@ieee.org

Yunsheng Ma
University of Massachusetts Medical School, Department of Medicine
55 Lake Avenue North, Worcester, MA 01655, USA
yunsheng.ma@umassmed.edu

ABSTRACT
As technology develops and research environment improves, large volume of data is collected for analyses. Unfortunately, these data are collected but not fully used or untouched. Particularly, such big data from health and medical studies pose significant challenges to the methodological field. This paper presents a new multi-clustering approach for pattern recognition of big data in a randomized controlled trial (RCT) with multi-validation criteria. Specifically, a nutritional dataset was used to demonstrate our approach, which was generated from an NIH-funded RCT for patients with metabolic syndrome. The proposed approach includes a suite of emerging and popular clustering methods: probability-based Gaussian Mixture Model (GMM), Hidden Markov Random Fields(HMRFs), Self-Organizing Map (SOM)-based neural networks, K-means and Agglomerative Hierarchical method. Using our RCT data and multi-validation criteria, our approach identified a most sufficient set of nutritional variables and detected distinct dietary change patterns with a universal agreement among the proposed multi-methods. The trajectory patterns were then generated using the method with the most clustering accuracy which was cross-validated via simulation. These patterns generated new and finer results for outcomes of the RCT. While our approach demonstrated a more accurate and comprehensive clustering only for nutritional data in RCT, it can be generalized to big data in other research fields.

Categories and Subject Descriptors
G.4 [**Mathematics of Computing**]: Mathematical Software; J.3 [**Life and Medical Sciences**]: Health; I.5 [**Pattern Recognition**]: Clustering; I.6 [**Computing Methodologies**]: Model Development

BODYNETS 2013, September 30-October 02, Boston, United States

DOI 10.4108/icst.bodynets.2013.253690

General Terms
Algorithms, Measurement, Design, Theory

Keywords
Big Data, random controlled trial, pattern recognition, heterogeneity, simulation, dietary quality, nutritional datasets, Gaussian Mixture Model (GMM), Hidden Markov Random Fields (HMRFs), Self-Organizing Map-based neural networks (SOM), K-means, Agglomerative Hierarchical Clustering.

1. Introduction
For randomized controlled trials (RCTs), patients are randomly assigned to different conditions and ideally balanced in assigned groups. However, a perfect randomization is rarely achieved in reality. Therefore, data are collected on a large variety of variables to better understand the efficacy of designed trials and patients' differential responses. These data range from demographic, anthropometric, physiological to biomarkers and blood lipid domains, forming Big Data. Pattern recognition methods are needed to cluster patients into different groups where their responses to the trials are similar, and hence to explain more variability in their outcomes.

Existing pattern recognition methods have individual advantages and obvious drawbacks. [7][9][11][13][19] This paper proposes a new multi-clustering approach with multi-validation criteria to identify patterns with higher accuracy and comprehensive validation.

To test our approach, we used a nutritional dataset from a two-arm physician-blinded randomized controlled trial. The overall goal of this trial is to compare the efficacy of two intervention approaches to dietary change among patients with metabolic syndrome. The two approaches are: 1) the American Heart Association (AHA) Dietary Guidelines, which is the current recommendation for patients with the metabolic syndrome and 2) a single dietary change condition that focuses exclusively on increasing fiber. For the "AHA condition," patients will be instructed to make dietary changes based on recommendations from the 2006AHA

guidelines. For the "high fiber diet condition," patients will be given specific instructions to increase sources of dietary fiber to a goal of ≥30 g per day, without specific instructions on other parts of their diet. Both conditions include an identical number of individual and group sessions led by dietitians. Three 24-hour dietary recalls were collected at baseline and at 3-, 6-, and 12-months after randomization. Women consist of 71.67% of the sample; the average age is 51.98 years old. About 88% of the patients are Whites, 2.5% Blacks, and 1.3% are Asians; 48.43% of patients have less than four years of college education.

The enrolled 240 patients responded to eight dietary quality components: 1) fruit, 2) vegetables, 3) nuts and legumes, 4) ratio of white to red meat, 5) cereal fiber, 6) trans-fat, 7) ratio of polyunsaturated fat to saturated fat, 8) alcohol. [1][2] Each component has four repeated measures. In addition to raw scores on these components, Alternative Healthy Eating Index (AHEI) scores were calculated to evaluate these components of a healthy cardiovascular diet [3]. Besides these nutritional variables, this RCT collected more than a hundred of other variables including repeated measures, such as anthropometrics, blood pressures, fasting blood glucose, glycosylated hemoglobin (HbA1c), blood lipid profiles, insulin, inflammatory markers, medication use, elevated depressive syndrome, quality of life, and levels of physical activity.

Our proposed multi-clustering approach was applied to this nutritional data set and identified dietary change patterns over the study period. As an example, we will only relate identified patterns to patients' weight changes, the primary outcome of this RCT, to demonstrate the utility of our approach. Our multi-clustering algorithm includes those from probabilistic based Gaussian Mixture Model (GMM), Hidden Markov Random FieldsHMRFs), Self-Organizing Map (SOM) based neural network, K-means and Hierarchical clustering. Our multi-validation criteria integrate Bayesian Information Criteria (BIC) and Deviance Information Criteria (DIC). We hypothesized our approach will provide more valid and accurate clustering and help explain more heterogeneity of trial effects on the outcomes of patients with metabolic syndrome.

This paper is organized as follows: Section 2 discusses related work on pattern recognition of big data in RCTS and typical clustering methods; Section 3 presents our proposed multi-clustering methods and multi-validation criteria; Section 4 illustrates the utility of our approach for the nutritional data of the RCT; and Section 5 concludes the findings of our research, limitations and future studies.

2. Related Work

Current RCTs often collect large data to better understand patients' response to trials [1][2]. Pattern recognition is a method to disentangle the complexity of patients' response and help clarify the efficacy of RCTs. Although not for RCTs, this technique was proposed as early as 1980's to extract information from big data [4]. The authors emphasized the importance of exacting information in an automatic way, and commented the advantages and drawbacks of each individual method for chemical and physical fields. More recently, mixture models were becoming popular; some researchers used this technique for observational studies. For example, Espy et al. [5] used this method to examine the detrimental effects of pregnant smoking on infants' neuropsychological development. Others used this method for identifying sucking patterns of breast-fed infants to understand to what extent variation is from infant to infant and from feed to feed [6].

Mixture models are built on Gaussian Mixture Model (GMM). It is a model-based approach assuming normal distributions for clusters. Liu L. and Yu Z. [7] proposed a practical computational method using Gaussian quadrature technique to obtain maximum likelihood estimates (MLE) for mixed models with non-normal random effects. This method is similar to traditional GMM based method with Expectation Maximization (EM) algorithm. Martella F. and Vermunt J.K. [8] proposed Gaussian based mixture of random effect models to cluster data from sibling pairs. In Martella's paper, a hierarchical mixture model with non-parametric random effects was proposed to capture the hierarchical structure of subjects, and a mixture of linear mixed effect models (LMMs) was used to cluster repeated gene expression data. However, all these models are parametric and assume statistical distributions and need subjective parameter adjustment in clustering.

Antonello et al.[9] proposed a finite mixture of non-homogeneous Non-Homogeneous Hidden Markov Models (NH-HMMs) to tackle the heterogeneity problem. This model allows them to consider observed sources of heterogeneity by means of a proper set of covariates, time and individual dependent. Their model is a finite mixture of NH-HMMs that can be used to classify individuals according to their dynamic behavior and to estimate a mixed NH-HMMs without any assumption regarding the distribution of the random term. Francois et al. [10] proposed a hierarchical Bayes clustering algorithm which incorporates models for geographical continuity of allele frequencies using hidden Markov random fields (HMRFs) as prior distributions. They assume Markov Chain Monte Carlo procedure can implement their algorithm efficiently and detect significant geographical discontinuities in allele frequencies and regulate the number of clusters. However, HN-HMMs lack criteria to appropriately identify the model with the best trade-off between fit and complexity. [9] The starting point for HN-HMMs model cannot be completely at random and it may lead to local maxima and degenerate solutions. Though HMRFs method has its advantages of accuracy in gene studies as authors evaluated, this method is based on hierarchical clustering, which has well-known static weakness. [10]

There is another popular clustering method based on neural networks. Vesanto et al. [11] proposed Self-Organizing Map (SOM) clustering with several approaches. Their approach was found to perform well when compared with direct clustering (e.g., Hierarchical and K-means clustering) and reduce computational time. SOM was applied to climate and geographical studies and was found to be a valuable tool for extracting characteristic surface current patterns. [12] However, SOM is unstable and systematically inaccurate in many cases, in addition to its computational cost. Additionally, this method cannot handle missing data.

K-means clustering dates back to 1967 and is one of the simplest unsupervised and effective learning algorithms that solve the well-known clustering problem [13]. This clustering has been widely used in image analyses. For example, Ng et al. [14] proposed this method for image segmentation. They use this clustering to obtain a primary segmentation of magnetic resonance images before applying their watershed segmentation. However, K-means does

not have criterion to select clusters and cannot handle missing values.

The hierarchical clustering is another commonly used clustering method. Researchers [15], [16], [17] [18] have presented survey and overview of hierarchical clustering. Matsui et al.[19] proposed a hierarchical clustering method to select disjoint gene clusters which have strong marginal association with disease related survival outcome from significant genes. McClelland et al.[20] proposed a regression based variable clustering method based on hierarchical clustering, to summarize the disease rates and clusters original regions into a reduced set of bigger areas. This method can tolerate moderately high level noise and performance was improved with increasing sample sizes. Hierarchical clustering seems to be widely used in gene studies. However, in addition to its inability to deal with missing values, hierarchical clustering is not able to incorporate information about the shape and size of clusters and has its static nature mentioned above.

Although popular in certain areas, all these methods described above seem to have their own drawbacks. To our understanding, our proposed multi-clustering method is the first to be used in RCTs, particularly in nutritional data for patients with metabolic syndrome.

3. Methods

The proposed multi-clustering methods with multi-validation steps are aimed to provide a more accurate and comprehensive pattern recognition of big data in RCTs. Specifically, our big nutritional datasets include over a hundred of variables for 240 patients with metabolic syndrome. We identified the most important dietary variables and patients' dietary trajectory patterns. Consequently, these patterns explained the heterogeneity of patients' dietary intake, and also facilitated interpreting the degrees of changes on anthropometric (e.g., weight, BMI, waist circumference), physiologic (e.g., fasting glucose, insulin, HbA1c) and blood lipid profile (e.g., HDL, LDL) outcomes.

3.1 Multi-Clustering approach for big RCT data

The framework of our multiple-clustering approach is introduced below. The rationale for using this approach is that no one knows the distribution of real data and which methods will generate the best clustering. Our approach provides a comparative validation of temerging and typical methods and identifies the patterns with the most accuracy rate, subsequently more trustworthy evaluation of trial effects on outcomes. For this big RCT data, we conducted stepwise validation to identify the optimal set of variables for pattern recognition of our big dataset. Our multi-clustering algorithm is integrated based on the following methods.

3.1.1 Gaussian Mixture Model (GMM) Clustering

We used the typical Expectation Maximization (EM) algorithm to estimate GMM parameters. The GMM is a weighted sum of M component Gaussian densities, expressed as follows: [21]

$$p(x|\lambda) = \sum_{i=1}^{M} \omega_i g(x|\mu_i, \sum_i) \quad (1)$$

where x denotes the features or attributes of big data such as in our nutritional data, λ is the initial model parameterized by mixture weights, mean vectors and covariance matrices: $\lambda = \{w_i, \mu_i, \Sigma_i\} \quad i = 1,...,M$, ω_i are the mixture weights from our datasets and $g(x|\mu_i, \sum_i)$ are the component Gaussian densities with μ_i as means of each component of our datasets and $\sum_i$ as covariance matrix from our datasets;

EM algorithm [22] is used to divide our RCT data, i.e. attributes/features, $X = \{x_1, x_2, ..., x_n\}$, into each cluster, $C = \{c_1, c_2, ..., c_m\}$. It starts with an estimation of unknown parameters and performs E- and M-steps iteratively:

E (Expectation) step is to estimate expected values of unknown parameters based on known parameters, expressed as;

$$p(k|n) = \frac{N(x_n|\mu_k, \Sigma_k)P(k)}{P(x_n)} \quad (2)$$

where the function $P(k)$ corresponds to the probability density function (PDF) of a given nutritional dataset x_n and $P(x_n)$ is the model probability at x_n and μ_k denotes the means and Σ_k denotes variance-covariance matrices of a specific cluster *c* within the mixture of M component under the constraint $\sum_{k=1}^{M} w_k = 1$. The PDF values assign membership of a specific input to a specific cluster. [22]

M (Maximization) step is to re-estimate the hidden parameters in order to maximize the likelihood of data, expressed as:

$$\text{M step: } \hat{\mu}_k = \frac{\sum_n p_{nk} x_n}{\sum_n p_{nk}} \quad (3)$$

$$\hat{\Sigma}_k = \frac{\sum_n p_{nk} \| (x_n - \hat{\mu}_k) \|^2}{\sum_n p_{nk}} \quad (4)$$

$$\hat{P}(k) = \frac{1}{N} \sum_n p_{nk} \quad (5)$$

In equations (3), (4) and (5), the means, variances and mixing weights for each cluster is updated respectively with the parameters calculated in equation (2). The E- and M- steps perform iteratively until the end of the algorithm. The algorithm termination condition is to either reach a satisfactory objective value or execution of a maximum number of iterations.

3.1.2 Hidden Markov Random Fields (HMRFs) Clustering

Hidden Markov Models (HMM) were first introduced in 1980s [23] as a tool for speech recognition. There are different types of HMM based clustering. For our multi-clustering approach we

integrated a method called Hidden Markov Random Field (HMRFs) [10]. HMRFs was first applied to gene studies, their approach is based on a hierarchical Markov Random Field (MRF) model whose neighborhood system is obtained from the Voronoi tessellation. HMRFs derives from Bayesian concepts where the prior distribution on cluster labels is defined as a HMRFs on a spatial individual network.

Here, according to this model, for our nutritional data, we denote input datasets as s_i and each s_i is surrounded by observations of participants which are closer to s_i than to any other observations. This set of points is known as the Dirichlet cell or tile. Two observations of participants are neighbors if their corresponding variables share a common edge. We denote c_i as the cluster from which the individual i originates and assume the existence of at most $K_{\max}$ clusters. The priors for our participants' datasets are Dirichlet distributions $D(\lambda,...,\lambda)$. The prior distribution for the set of cluster configurations is defined as a Gibbs distribution:

$$\pi(c) = \exp\left[\psi U(c)\right] / Z,\ c \in \{1,...,K_{\max}\}^N \tag{6}$$

where ψ is a non-negative parameter called the interaction parameter, $U(c)$ is the number of neighboring pairs that share the same labels in c, and Z is a normalizing constant called the partition function. Inferences on (c, f) are carried out by simulating the posterior distribution $\pi(c, f \mid z)$ through a Markov Chain Monte Carlo (MCMC) sampling algorithm.

3.1.3 K-means Clustering

K-means is one of the simplest but effective unsupervised learning algorithms for clustering. It aims to minimize the Euclidean distance between the data points and the corresponding cluster centroid, which is achieved by minimizing the objective function:

$$\sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^j - c_j\right\|^2 \tag{7}$$

where x_i^j is our RCT nutritional dataset, and c_j is the cluster center. The clustering steps are as follows:

Step 1: Define K centroids, one for each cluster.

Step 2: After placing centroids, associate each point (the observation on each component of our nutritional datasets) to the nearest centroid. When no point is pending, the first step is completed and the first initial clusters were generated.

Step 3: Re-calculate K new centroids as centers of the clusters resulting from the previous step and a new binding will be completed between the same dataset points and the nearest new centroid.

A loop has been generated through step1 to step 3. As a result of this loop, the K centroids change their locations step by step till no more changes.

3.1.4 Self-Organizing Map (SOM) Clustering

The Self-Organizing Map (SOM) is a type of artificial neural network trained using an unsupervised and competitive learning process to produce a low-dimensional model. The SOM was developed by Kohonen [24], where the updating of weights can be modified to involve neighboring relations in the output array.

To apply this method for our nutritional data, we define R^n as a vector space, and $X\ (x_1,...,x_n)$ as input data which are our nutritional datasets. Let $x(t)$ be our input datasets at step t and let $w_i(0)$ be weight vectors at initial values in R^n space. For given input vector $x(t)$, we calculate the distance between $x(t)$ and $w_i(t)$, and select the weight vector as winner ψ minimizing the distance. The process is to seek:

$$\psi = \arg\min\left\{\left\|x - w_i\right\|\right\} \tag{8}$$

where $\arg(.)$ generates index ψ of the winner. With the use of winner ψ, weight vector $w_i(t)$ is updated:

$$w_i = \begin{cases} \alpha(t)(x - w_i) & (i \in N_c(t)) \\ 0 & (otherwise) \end{cases} \tag{9}$$

where $\alpha(t)$ is the learning rate and is a decreasing function of time. $N_c(t)$ has a set of indices of topological neighborhoods for winner ψ at step t.

The adaptive learning algorithm evaluates unknown probability density function $p(x)$. Weight vectors represent centroids of each clustering set. Cost function H is expressed as:

$$H = \sum_{i=1}^{k}\int_{S_i} d(x, w_i)\, p(x)\, dx \tag{10}$$

where k is the number of clustering set represented by partition space S_i and $d(x, w_i)$ is the square error of the Euclidean distance between attribute/feature vector $X = (x_1,...,x_n)$ and weight vector $w_i = (w_{1i},...,w_{ni})$.

3.1.5 Hierarchical Clustering

Here we focus on agglomerative hierarchical clustering, also called bottom-up clustering, where each observation starts in its own cluster and merges as one moves up the hierarchy. The clustering steps are as follows:

Step 1: Start by assigning each observation into each cluster. For our 240 patients we first treat them as 240 clusters, and assume the distance or similarity between clusters is the same as those between observations they contain.

Step 2: Find the closest or most similar pair of clusters and merge them into one cluster.

Step 3: Compute the distance or similarity between the new cluster and previous clusters.

Step 4: Repeat Step 2 and 3 until all observations are clustered into pre-defined N clusters.

The choice of different distance metric will influence the results of hierarchical clustering. To illustrate the process, we use the common Euclidean distance:

$$d(a,b) = \left\|a - b\right\|_2 = \sqrt{\sum_i (a_i - b_i)^2} \tag{11}$$

Where $d(a,b)$ is our chosen Euclidean distance metric, a and b are two random observations for calculating distance or similarity.

3.2 Criteria for Clustering

To identify the optimal number of clusters, we propose to use multi-validation criteria. The Bayesian Information Criterion (BIC) was used for GMM K-means and SOM, while Deviance Information Criterion (DIC) was used for typical Bayesian-based method, Hidden Markov Random Fields(HMRFs).

3.2.1 Bayesian Information Criterion (BIC)

The Bayesian Information Criterion (BIC) is for model selection among a finite set of models. It is based on the likelihood function, expressed as:

$$BIC = -2 \cdot \ln p(x|k) = -2 \cdot \ln L + k \ln(n) = n \cdot \ln(\hat{\sigma}_e^2) + k \cdot \ln(n) \quad (12)$$

where x is denoted the same as above; n is the number of data points in x, i.e., sample size; k is the number of free parameters to be estimated; $p(x|k)$ is the likelihood of the parameters given our nutritional data features, L is the maximized value of the likelihood function for the estimated model, $\hat{\sigma}_e^2$ is the error variance, defined as:

$$\hat{\sigma}_e^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2 \quad (13)$$

Lower BIC indicates better clustering. The BIC is an increasing function of σ_e^2 and k, so the lower BIC implies either fewer explanatory variables, better fit or both. The BIC resolves the over-fitting problem by introducing a penalty term for the number of parameters in the model.

3.2.2 Deviance Information Criteria (DIC)

The Deviance Information Criterion (DIC) is a hierarchical modeling generalization of the BIC [10]. It is a measure of model complexity computed as the model deviance penalized by an estimate of the effective number of parameters. It is particularly useful in Bayesian model selection where the posterior distributions of the models are obtained by Markov chain Monte Carlo (MCMC) simulation. In other words, the DIC is well adapted to MCMC simulations, and smaller DIC indicates better clustering. Assume that θ is a vector containing all the algorithm parameters. Taking our data as example, $D(\theta)$ is denoted as the model deviance for the parameter set θ computed as -2 times the log likelihood generated by HMRFs, and p_D the effective number of parameters in the model. The DIC can be calculated as:

$$DIC = \overline{D} + p_D \quad (14)$$

where $\overline{D}$ denotes the posterior mean of the deviance, defined as: $\overline{D} = E^{\theta}[D(\theta)]$. It is a measure of how well the model fits the data, where $D(\theta) = -2\log(p(x|\theta)) + \mathrm{A}$, x is our data features, θ are the unknown parameters of HMRFs and $p(x|\theta)$ is the likelihood function. Symbol A is an unknown constant that cancels out in all calculations that compare different models.

4. Empirical Results with Simulation

Our multiple-clustering approach generates an integrated algorithm by including the above five clustering methods, and includes multi-validation indices for validating optimal cluster numbers. Our approach unanimously point to 3 clusters for this big nutritional data and K-means has the highest accuracy rate for this RCT study. Additionally, our approach help identify optimal set of variables for clustering these nutritional data and new findings were detected for the study outcome.

4.1 Multi-Validation Criteria to Choose Optimal Cluster Numbers

BIC was used to test GMM, K-means, Hierarchical and SOM clustering methods and DIC was used to test Hidden Markov Random Fields (HMRFs). Smaller BIC and DIC values indicate better clustering.

Table 1. Multi-validation indices for cluster selection

	2 cls	**3 cls**	**4 cls**	**5 cls**	**6 cls**	**7 cls**	**8 cls**
GMM	34220	**14309**	14767	16064	20653	25566	33233
K-m	40406	**12315**	16690	17494	22745	25755	36766
SOM	1220	**578**	1825	2815	2228	2885	2998
Hierar chical	5375	**3765**	3985	4158	5778	8196	9263
HMRF	13.93	**5.96**	14.72	17.06	18.68	20.16	21.81

The lowest DIC and BIC values indicate that our multi-clustering methods identify three optimal clusters for this big nutritional data set.

4.2 Cross-Validate Clustering Accuracy via simulation

To assess the clustering accuracy of each method, we used the parameters of our nutritional data and simulated 3 clusters given the parameters from our nutritional clusters. The accuracy rates are listed in Table 2.

Table 2. Accuracy rate for each clustering method

	K-means	**GMM**	**HMRF**	**SOM**	**Hierarhcial**
Accuracy Rate	91.33%	85.32%	75.52%	78.89%	61.77%

K-means showed the highest accuracy, followed by GMM, HMRF, and SOM. Hierarchical method has the lowest accuracy as shown in our previous research. [5][6]

4.3 Indentify the optimal set of variables for big nutritional data

We used stepwise method to find the most accurate and parsimonious set of nutritional variables for clustering:

Step 1: Patients' scores on 8 nutritional components of American Heart Association's Eating Index (AHEI): fruit, vegetables, nuts and legumes, ratio of white to red meat, cereal fiber, trans-fat, ratio of polyunsaturated fat to saturated fat, and alcohol (8*4 time points = 32 variables);
Step 2: Patients' AHEI scores on 8 nutritional components and total AHEI scores (9*4 time points = 36 variables);
Step 3: Patients' raw values on 8 nutritional components (8*4 time points = 32 variables);
Step 4: Step 2 and Step 3 variables (68 variables).

Our results show that the set of variables for Step 3 generated the lowest but similar clustering accuracy rate and therefore were identified as optimal.

4.4 Statistics and Graphs for each cluster

The identified clusters show that patients displayed three dietary intake patterns for this RCT study (Figure 1). Although with fluctuation in their intake, patients in Cluster 1 seem to have the best dietary behaviors, with higher dietary trajectories (or scores) than the other two clusters across most (7 out of 8) dietary components (Figure 1) and the highest dietary quality score (i.e., AHEI total scores) . Compared to Cluster 1 patients, Cluster 2 and 3 on average have lower dietary score trajectories, with Cluster 3 as the lowest across 7 dietary components (Figure 1).

In terms of weight loss, patients in all three clusters on average, maintained constant weight loss over the study period. Cluster 1 ranks the highest for baseline weight (220.92 lbs), followed by Cluster 2 (219.95 lbs) and Cluster 3 (215.66 lbs.). Corresponding to the order of their dietary score patterns, Cluster 1 patients, on average, rank the first in weight loss from baseline to 12 months (~22 lbs), with Cluster 2 (~ 20 lbs) and Cluster 3 (~19 lbs) as the second and third. This may indicate the degree of nutritional diet intake is related to the degree of weight loss for these patients with metabolic syndrome.

Moreover, for each cluster, the number of patients from each RCT condition, called intervention (high fiber diet condition) and control (AHA condition), are almost identically distributed across each cluster. This may imply that the intervention group with the a simple dietary message following the high-fiber guideline plays as well as the control group following more complicated AHA guideline in terms of dietary quality. Also, it shows that, regardless of the treatment conditions, the extent of the dietary quality (a proxy for different dietary behaviors) may contribute to the heterogeneity of trial effects on weight loss for patients with metabolic syndrome.

Table 4. Patients of intervention and control groups in each cluster

	Cluster1	Cluster 2	Cluster 3
Intervention	48.33%	49.27%	50.45%
Control	51.67%	50.73%	49.55%

5. Conclusions and Future Work

In this paper, we proposed a new multi-clustering approach with multi-validation criteria for big random control trials (RCT) data. We developed an integrated algorithm based on five emerging and typical clustering methods, Gaussian Mixture Model (GMM), Hidden Markov Random Field (HMRF), K-means, Self-Organizing Map-based (SOM) and Hierarchical Clustering. Bayesian Information Criterion (BIC) and Deviance information Criterion (DIC) were used to select the optimal number of clusters. For this working RCT example, our stepwise approach helped identify the optimal number of nutritional variable for clustering. Although all methods point to 3 clusters based on validation indices, K-means was found to be most accurate to generate the dietary change patterns for this big data. The identified trajectory patterns generated more interesting findings on the relationships among dietary behaviors, weight and trial conditions.

Our approach could be extended to analyze big data in other fields such as those generated by body sensors. This study has limitations for missing data imputation, as we only used single imputation for our RCT data. This RCT study was only used to demonstrate our approach and we only examined one primary outcome of weight loss, therefore, the conclusions for this RCT study may not be generalizable. In future studies, multiple imputation method will be considered for missing data and we will test our approach for other big data of RCT studies and more simulation work will be conducted.

6. ACKNOWLEDGMENTS

The project described was supported by the pilot project program awarded to Dr. Hua Fang from the National Center for Research Resources UL1RR031982 and partly by grant HR01HL094575 to Dr.Yunsheng Ma from the National Heart, Lung and Blood Institute (NHLBI).

7. REFERENCES

[1] Ma Y., Pagoto SL., Griffith JA., et al. A Dietary Quality Comparison of Popular Weight-Loss Plans. Journal of the American Dietetic Associtation 2007; 107:1786-91.

[2] Ma Y., Li W., Olendzki B, et al. Dietary Quality q year after diagnosis of coronary heart disease. J Am Diet Assoc 2008; 18:240-6.

[3] McCullough ML., Feskanich D., Stampfer MJ., Giovannucci EL., Rimm EB., Hu FB., Spiegelman D., Hunter DJ., Colditz GA., Willet WC. Diet quality and major chronic disease risk in men and women: moving toward improved dietary guidance. Am J Clin Nutr. 76(6):1261-1271, 2002 Add McCullough, 2006 #2994; McCullough, 2002

[4] M.P. Derde, D.L. Massart. Extraction of information from large data sets by pattern recognition. *Fresenius Zeitschrift fur analytische Chemiet.* 1982, Volume 313, Issue 6, pp 484-495.

[5] Espy, K. A., Fang, H., Charak, D., Minich, N., & Taylor, H. G. (2009). Growth Mixture Modeling of Academic Achievement in Children of Varying Birth Weight Risk. *Neuropsychology. 23(4), 460-474*.

[6] Chetwynd, A.G., Diggle, P.J., Drewett, R.F. and Young, B. (1998), A mixture model for sucking patterns of breast-fed infants. *Statist. Med.,* 17: 395-405.

[7] Liu, L. and Yu, Z. (2008), A likelihood reformulation method in non-normal random effects models. *Statist. Med.,* 27:3105-3124.

[8] Martella, F., Vermunt, J.K., Beekman, M., Westendorp, R.G.J, Slagboom, P.E. and Houwing-Duistermaat, J.J. (2011), A mixture model with random-effects components for classifying sibling pairs. *Statist. Med.,* 30:3252-3264.

[9] Maruotti, A. and Rocci, R. (2012), A mixed non-homogeneous hidden Markov model for categorical data,

with application to alcohol consumption. *Statist. Med.,* 31:871-886.

[10] O.Francois, S.Ancelet, G.Guillot. Bayesian clustering using Hidden Markov Random Fields in spatial population genetics. *Genetrics.,* 2006 October; 174(2):805-816.

[11] Vesanto J. and Alhoniemi E. (2000), Clustering of the self-organizing map. *IEEE Trans Neural Netw.,* 2000;11 (3):586-600.

[12] Mihanović, H., S. Cosoli, I. Vilibić, D. Ivanković, V. Dadić, *and* M. Gačić *(*2011*),* Surface current patterns in the northern Adriatic extracted from high-frequency radar data using self-organizing map analysis, J. Geophys. Res., 116, *C08033*

[13] J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability"*, Berkeley, University of California Press, 1:281-297

[14] H.P. Ng, S.H. Ong, K.W.C. Foong, P.S. Goh, W.L. Nowinski, Medical Image Segmentation Using K-Means Clustering and Improved Watershed Algorithm, *SSIAI*, pp.61-65, 2006 IEEE Southwest Symposium on Image Analysis and Interpretation, 2006

[15] A.K. Jain, M.N. Murty and P.J. Flynn. Data Clustering: A Review. *Journal of ACM Computing Surveys*. Vol.31, Issue 3, Sept.1999 pp. 264-323.

[16] A. D. Gordon. A Review of Hierarchical Classification. *Journal of the Royal Statistical Society*. Series A (General), Vol. 150, No. 2 (1987), pp. 119-137

[17] B. Mirkin. 1998. Mathematical Classification and Clustering: From How to What and Why. *Classification, Data Analysis and Data Highways, Studies in Classification, Data Analysis and Knowledge Organization*, 1998, pp. 172-181.

[18] A. Griffiths, L.A. Robinson and P. Willett, (1984) Hierarchic Agglomerative Clustering Methods for Automatic Document Classification. *Journal of Documentation*, Vol. 40 Iss: 3, pp.175 - 205

[19] Matsui, S., Yamanaka, T., Barlogie, B., Shaughnessy, J. D. and Crowley, J. (2008), Clustering of significant genes in prognostic studies with microarrays: Application to a clinical study for multiple myeloma. Statist. Med., 27: 1106–1120. doi: 10.1002/sim.2997

[20] McClelland, R. L. and Kronmal, R. A. (2002), Regression-based variable clustering for data reduction. Statist. Med., 21: 921–941. doi: 10.1002/sim.1063

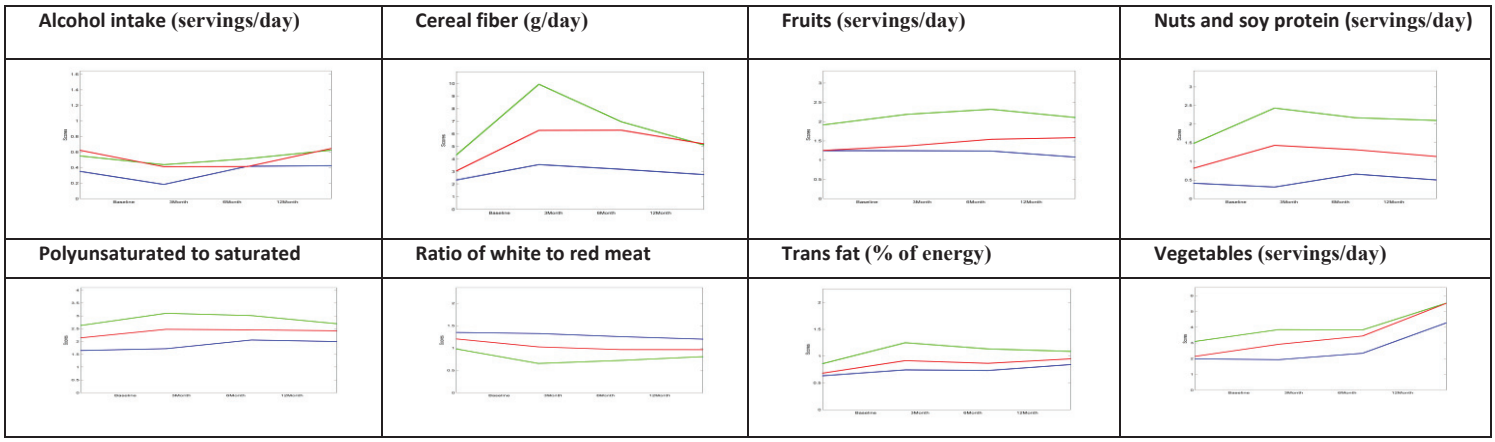
[21] Reynolds, D.A. (1992), A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification. PhD thesis, Georgia Institute of Technology (1992).

[22] Karabulut, M. and Ibrikci, T. (2012), Identification of transcription factor binding sites using Gaussian mixture models. Expert Systems. doi: 10.1111/exsy.12004

[23] Rabiner, L. R. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE 77,2 (Feb. 1989), 257-285

[24] T. Kohonen. Self-Organizing Maps. Berlin/Heidelberg, Germany: Springer, 1995, vol.3.

Figure 1. (a) mean dietary change patterns for each cluster (cluter1: Green line; cluster2: Red line; cluster3: Blue line)



(b) AHEI score for each cluster