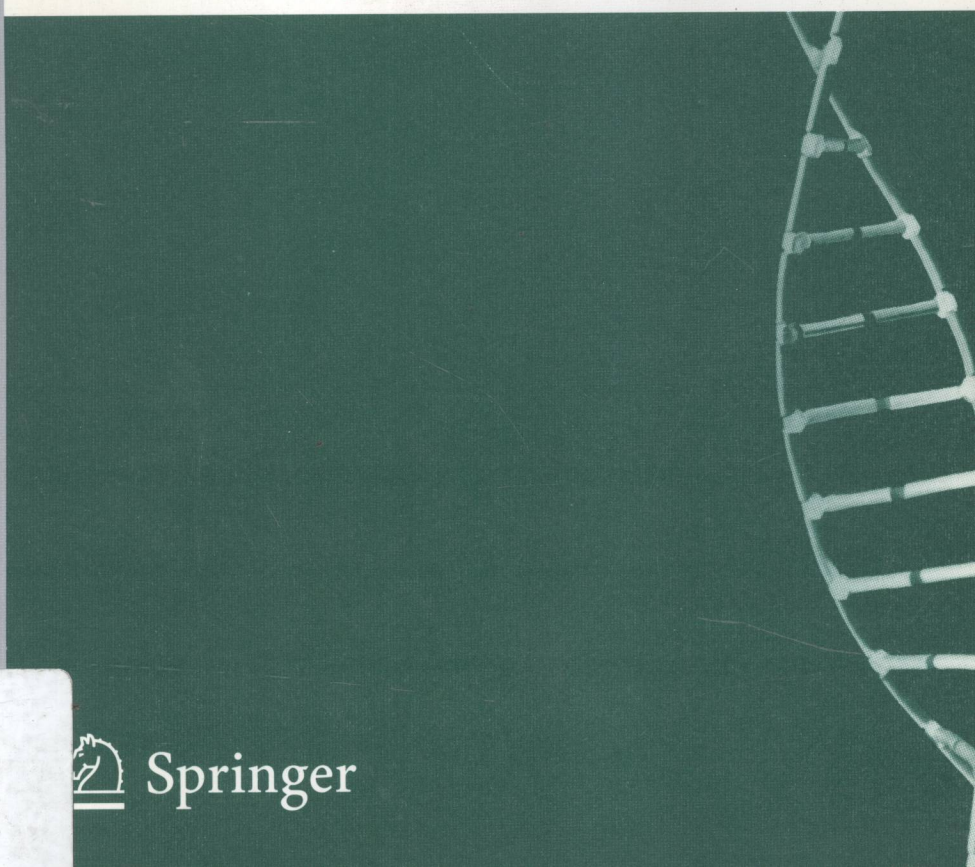
Mehmet M. Dalkilic
Sun Kim
Jiong Yang (Eds.)

LNBI 4316

Data Mining and Bioinformatics

**First International Workshop, VDMB 2006
Seoul, Korea, September 2006
Revised Selected Papers**



Springer

Mehmet M. Dalkilic Sun Kim
Jiong Yang (Eds.)

Data Mining and Bioinformatics

First International Workshop, VDMB 2006
Seoul, Korea, September 11, 2006
Revised Selected Papers

Springer

Series Editors

Sorin Istrail, Brown University, Providence, RI, USA
Pavel Pevzner, University of California, San Diego, CA, USA
Michael Waterman, University of Southern California, Los Angeles, CA, USA

Volume Editors

Mehmet M. Dalkilic
Sun Kim

Indiana University
Center for Genomics and Bioinformatics
School of Informatics
Bloomington,IN 47408, USA
E-mail: {dalkilic,sunkim2}@indiana.edu

Jiong Yang
Case Western Reserve University
EECS Department, Department of Epidemiology and Biostatistics
Cleveland, OH, 44106, USA
E-mail: jiong.yang@case.edu

Library of Congress Control Number: 2006938545

CR Subject Classification (1998): H.2.8, I.5, J.3, I.2, H.3, F.1-2

LNCS Sublibrary: SL 8 – Bioinformatics

ISSN 0302-9743
ISBN-10 3-540-68970-2 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-68970-6 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11960669 06/3142 5 4 3 2 1 0

Lecture Notes in Bioinformatics 4316

Edited by S. Istrail, P. Pevzner, and M. Waterman

Subseries of Lecture Notes in Computer Science

Lecture Notes in Bioinformatics

Vol. 4345: N. Maglaveras, I. Chouvarda, V. Koutkias, R. Brause (Eds.), Biological and Medical Data Analysis. XIII, 496 pages. 2006.

Vol. 4316: M.M. Dalkilic, S. Kim, J. Yang (Eds.), Data Mining and Bioinformatics. VIII, 197 pages. 2006.

Vol. 4230: C. Priami, A. Ingólfsdóttir, B. Mishra, H.R. Nielson (Eds.), Transactions on Computational Systems Biology VII. VII, 185 pages. 2006.

Vol. 4220: C. Priami, G. Plotkin (Eds.), Transactions on Computational Systems Biology VI. VII, 247 pages. 2006.

Vol. 4216: M.R. Berthold, R. Glen, I. Fischer (Eds.), Computational Life Sciences II. XIII, 269 pages. 2006.

Vol. 4210: C. Priami (Ed.), Computational Methods in Systems Biology. X, 323 pages. 2006.

Vol. 4205: G. Bourque, N. El-Mabrouk (Eds.), Comparative Genomics. X, 231 pages. 2006.

Vol. 4175: P. Bücher, B.M.E. Moret (Eds.), Algorithms in Bioinformatics. XII, 402 pages. 2006.

Vol. 4146: J.C. Rajapakse, L. Wong, R. Acharya (Eds.), Pattern Recognition in Bioinformatics. XIV, 186 pages. 2006.

Vol. 4115: D.-S. Huang, K. Li, G.W. Irwin (Eds.), Computational Intelligence and Bioinformatics, Part III. XXI, 803 pages. 2006.

Vol. 4075: U. Leser, F. Naumann, B. Eckman (Eds.), Data Integration in the Life Sciences. XI, 298 pages. 2006.

Vol. 4070: C. Priami, X. Hu, Y. Pan, T.Y. Lin (Eds.), Transactions on Computational Systems Biology V. IX, 129 pages. 2006.

Vol. 3939: C. Priami, L. Cardelli, S. Emmott (Eds.), Transactions on Computational Systems Biology IV. VII, 141 pages. 2006.

Vol. 3916: J. Li, Q. Yang, A.-H. Tan (Eds.), Data Mining for Biomedical Applications. VIII, 155 pages. 2006.

Vol. 3909: A. Apostolico, C. Guerra, S. Istrail, P. Pevzner, M. Waterman (Eds.), Research in Computational Molecular Biology. XVII, 612 pages. 2006.

Vol. 3886: E.G. Bremer, J. Hakenberg, E.-H.(S.) Han, D. Berrar, W. Dubitzky (Eds.), Knowledge Discovery in Life Science Literature. XIV, 147 pages. 2006.

Vol. 3745: J.L. Oliveira, V. Maojo, F. Martín-Sánchez, A.S. Pereira (Eds.), Biological and Medical Data Analysis. XII, 422 pages. 2005.

Vol. 3737: C. Priami, E. Merelli, P. Gonzalez, A. Omicini (Eds.), Transactions on Computational Systems Biology III. VII, 169 pages. 2005.

Vol. 3695: M.R. Berthold, R. Glen, K. Diederichs, O. Kohlbacher, I. Fischer (Eds.), Computational Life Sciences. XI, 277 pages. 2005.

Vol. 3692: R. Casadio, G. Myers (Eds.), Algorithms in Bioinformatics. X, 436 pages. 2005.

Vol. 3680: C. Priami, A. Zelikovsky (Eds.), Transactions on Computational Systems Biology II. IX, 153 pages. 2005.

Vol. 3678: A. McLysaght, D.H. Huson (Eds.), Comparative Genomics. VIII, 167 pages. 2005.

Vol. 3615: B. Ludäscher, L. Raschid (Eds.), Data Integration in the Life Sciences. XII, 344 pages. 2005.

Vol. 3594: J.C. Setubal, S. Verjovski-Almeida (Eds.), Advances in Bioinformatics and Computational Biology. XIV, 258 pages. 2005.

Vol. 3500: S. Miyano, J. Mesirov, S. Kasif, S. Istrail, P. Pevzner, M. Waterman (Eds.), Research in Computational Molecular Biology. XVII, 632 pages. 2005.

Vol. 3388: J. Lagergren (Ed.), Comparative Genomics. VII, 133 pages. 2005.

Vol. 3380: C. Priami (Ed.), Transactions on Computational Systems Biology I. IX, 111 pages. 2005.

Vol. 3370: A. Konagaya, K. Satou (Eds.), Grid Computing in Life Science. X, 188 pages. 2005.

Vol. 3318: E. Eskin, C. Workman (Eds.), Regulatory Genomics. VII, 115 pages. 2005.

Vol. 3240: I. Jonassen, J. Kim (Eds.), Algorithms in Bioinformatics. IX, 476 pages. 2004.

Vol. 3082: V. Danos, V. Schachter (Eds.), Computational Methods in Systems Biology. IX, 280 pages. 2005.

Vol. 2994: E. Rahm (Ed.), Data Integration in the Life Sciences. X, 221 pages. 2004.

Vol. 2983: S. Istrail, M.S. Waterman, A. Clark (Eds.), Computational Methods for SNPs and Haplotype Inference. IX, 153 pages. 2004.

Vol. 2812: G. Benson, R.D.M. Page (Eds.), Algorithms in Bioinformatics. X, 528 pages. 2003.

Vol. 2666: C. Guerra, S. Istrail (Eds.), Mathematical Methods for Protein Structure Analysis and Design. XI, 157 pages. 2003.

Preface

This volume contains the papers presented at the inaugural workshop on Data Mining and Bioinformatics at the 32nd International Conference on Very Large Data Bases (VLDB). The purpose of this workshop was to begin bringing together researchers from database, data mining, and bioinformatics areas to help leverage respective successes in each to the others. We also hope to expose the richness, complexity, and challenges in this area that involves mining very large complex biological data that will only grow in size and complexity as genome-scale high-throughput techniques become more routine. The problems are sufficiently different enough from traditional data mining problems (outside of life sciences) that novel approaches must be taken to data mine in this area. The workshop was held in Seoul, Korea, on September 11, 2006.

We received 30 submissions in response to the call for papers. Each submission was assigned to at least three members of the Program Committee. The Program Committee discussed the submission electronically, judging them on their importance, originality, clarity, relevance, and appropriateness to the expected audience. The Program Committee selected 15 papers for presentation. These papers are in the areas of microarray data analysis, bioinformatics system and text retrieval, application of gene expression data, and sequence analysis. Because of the format of the workshop and the high number of submissions, many good papers could not be included. Complementing the contributed papers, the program of VDMB 2006 included an invited talk by Simon Mercer, Program Manager for External Research, with an empahsis on life sciences.

We would like to thank the members of the Program Committee for their hard and expert work. We would also like to thank the VLDB organizers, the external reviewers, the authors, and the participants for their contribution to the continuing success of the workshop. Thanks also to Indiana University School of Informatics for the generous financial support.

October 2006

Mehmet Dalkilic
Sun Kim
Jiong Yang
Program Chairs
VDMB 2006

VDMB 2006 Organization

Program Committee Chairs

Mehmet Dalkilic (Indiana University, USA)
Sun Kim (Indiana University, USA)
Jiong Yang (Indiana University, USA)

Program Committee Members

Mark Adams (Case Western University, USA)
Xue-wen Chen (University of Kansas, USA)
Jong Bhak (Korea Bioinformatics Center, Korea)
Dan Fay (Microsoft/Director Technical Computing, North America)
Hwan-Gue Cho (Busan National University, Korea)
Jeong-Hyeon Choi (Indiana University, USA)
Tony Hu (Drexel University, USA)
Jaewoo Kang (North Carolina State University, USA)
George Karypis (University of Minnesota, USA)
Doheon Lee (KAIST, Korea)
Jing Li (Case Western Reserve University, USA)
Yanda Li (Tsinghua University, China)
Birong Liao (Eli Lilly, USA)
Li Liao (University of Delaware, USA)
Huiqing Liu (University of Georgia, USA)
Lei Liu (University of Illinois at Urbana Champaign, USA)
Xiangjun (Frank) Liu (Tsinghua University, China)
Qingming Luo (Huazhong University, China)
Simon Mercer (Microsoft, USA)
Jian Pei (Simon Fraser University, Canada)
Meral Ozsoyoglu (Case Western Reserve University, USA)
Predrag Radivojac (Indiana University, USA)
Tetsuo Shibuya (University of Tokyo, Japan)
Keiji Takamoto (Case Western Reserve University, USA)
Haixu Tang (Indiana University, USA)
Anthony Tung (National University of Singapore, Singapore)
Wei Wang (University of North Carolina at Chapel Hill, USA)
Mohammed Zaki (Rensselaer Polytechnic Institute, USA)
Aidong Zhang (State University of New York at Buffalo, USA)

Table of Contents

Bioinformatics at Microsoft Research

Simon Mercer

Microsoft Research, One Microsoft Way, Redmond, WA 98052-6399
simon.mercer@microsoft.com

Abstract. The advancement of the life sciences in the last twenty years has been the story of increasing integration of computing with scientific research, and this trend is set to transform the practice of science in our lifetimes. Conversely, biological systems are a rich source of ideas that will transform the future of computing.

In addition to supporting academic research in the life sciences, Microsoft Research is a source of tools and technologies well suited to the needs of basic scientific research. Current projects include new languages to simplify data extraction and processing, tools for scientific workflows, and biological visualization.

Computer science researchers also bring new perspectives to problems in biology, such as the use of schema-matching techniques in merging ontologies, machine learning in vaccine design, and process algebra in understanding metabolic pathways.

M.M. Dalkilic, S. Kim, and J. Yang (Eds.): VDMB 2006, LNBI 4316, p. 1, 2006.

A Novel Approach for Effective Learning of Cluster Structures with Biological Data Applications

Miyoung Shin

School of Electrical Engineering and Computer Science, Kyungpook National University,
1370 Sankyuk-dong, Buk-gu, Daegu 702-701, Korea
shinmy@knu.ac.kr

Abstract. Recently DNA microarray gene expression studies have been actively performed for mining unknown biological knowledge hidden under a large volume of gene expression data in a systematic way. In particular, the problem of finding groups of co-expressed genes or samples has been largely investigated due to its usefulness in characterizing unknown gene functions or performing more sophisticated tasks, such as modeling biological pathways. Nevertheless, there are still some difficulties in practice to identify good clusters since many clustering methods require user's arbitrary selection of the number of target clusters. In this paper we propose a novel approach to systematically identifying good candidates of cluster numbers so that we can minimize the arbitrariness in cluster generation. Our experimental results on both synthetic dataset and real gene expression dataset show the applicability and usefulness of this approach in microarray data mining.

1 Introduction

In recent years, microarray gene expression studies have been actively pursued for mining biologically significant knowledge hidden under a large volume of gene expression data accumulated by DNA microarray experiments. Particularly great attentions have been paid to data mining schemes for gene function discovery, disease diagnosis, regulatory network inference, pharmaceutical target identification, etc [1, 2, 3]. A principal task in investigating there problems is to identify gene groups or samples which show similar expression patterns over multiple experimental conditions. The detection of such co-expressed genes or samples allows us to infer their high possibility to have similar biological behaviors, so these can be used to characterize unknown biological facts as in [4,5,6,7,8,9].

So far, numerous methods have been investigated to efficiently find groups of genes or samples showing similar expression patterns. An extensive survey of clustering algorithms is given in [10]. The most widely-used algorithms for microarray data analysis are hierarchical clustering [4], *k*-means clustering [8], self-organizing maps [5], etc. Also, there are more sophisticated algorithms such as quantum clustering

M.M. Dalkilic, S. Kim, and J. Yang (Eds.): VDMB 2006, LNBI 4316, pp. 2 – 13, 2006.

with singular value decomposition [11], bagged clustering [12], diametrical clustering [13], CLICK [14], and so on.

Nevertheless, there still remain some difficulties in practice to identify good clusters in an efficient way. One of the problems is that many clustering methods require user's arbitrary selection of the number of target clusters. Moreover, the selection of the number of clusters dominates the quality of clustering results. Some recent tasks have addressed these issues for cluster analysis of gene expression data. In [15], Bolshakova *et al.* estimated the number of clusters inherent in microarray data by using the combination of several clustering and validation algorithms. On the other hand, in [16], Amato *et al.* proposed an automatic procedure to get the number of clusters present in the data as a part of a multi-step data mining framework composed of a non-linear PCA neural network for feature extraction and probabilistic principal surfaces combined with an agglomerative approach based on Negentropy. Also, a recent paper by Tseng *et al.* [17] suggested a parameterless clustering method called the correlation search technique. Yet, these methods are either still based on an arbitrary selection of the number of clusters or work only with their own clustering algorithms.

In this paper our concern is to propose a systematic approach to identify *good* number of clusters on a given data, which also can possibly work with the widely-used clustering algorithms requiring a specified number k of clusters. To realize this, we define the goodness of the number of clusters in terms of its *representational capacity* and investigate its applicability and usability in learning cluster structures with synthetic dataset and microarray gene expression dataset. The rest of the paper is organized as follows. In Section 2, we give the definition of the *representational capacity* (hereafter *RC*) and introduce its properties. Based on these, in Section 3, we present the *RC*-based algorithm for the estimation of the number of clusters on a given dataset. In Section 4, the experimental results are presented and discussed. Finally, concluding remarks are given in Section 5.

2 Definition of *RC* and Its Properties

One of the critical issues in cluster analysis is to identify the good number of clusters on a given dataset. Intuitively it may be the number of groups in which the members within the group are highly homogeneous and the members between the groups are highly separable. Without a *priori* knowledge, however, it is not easy to conjecture the good number of clusters hidden under the data in advance. To handle this issue in an efficient and systematic way, we introduce the concept of *RC* as a vehicle to quantify the goodness of the cluster number and use this to estimate the good number of clusters for given data.

In this section, the definition of *RC* and its properties are given first, and then present the algorithm to estimate the number of clusters using *RC* criterion in the following section.

2.1 Distribution Matrix

To define the *RC*, we employ the matrix which captures the underlying characteristics of given data, called the *distribution matrix*. Specifically, for the dataset $\mathbf{D} = \{\mathbf{x}_i, i = 1,\ldots,n : \mathbf{x}_i = (x_{i1},\ldots,x_{id}) \in R^d\}$, the distribution matrix $\mathbf{\Phi}$ is defined as

$$\mathbf{\Phi} = \begin{bmatrix} \phi_{11} & \phi_{12} & \cdots & \phi_{1n} \\ \phi_{21} & \phi_{22} & \cdots & \phi_{2n} \\ \vdots & \vdots & & \vdots \\ \phi_{n1} & \phi_{n2} & \cdots & \phi_{nn} \end{bmatrix} = \begin{bmatrix} \phi(\mathbf{x}_1,\mathbf{x}_1) & \phi(\mathbf{x}_1,\mathbf{x}_2) & \cdots & \phi(\mathbf{x}_1,\mathbf{x}_n) \\ \phi(\mathbf{x}_2,\mathbf{x}_1) & \phi(\mathbf{x}_2,\mathbf{x}_2) & \cdots & \phi(\mathbf{x}_2,\mathbf{x}_n) \\ \vdots & \vdots & & \vdots \\ \phi(\mathbf{x}_n,\mathbf{x}_1) & \phi(\mathbf{x}_n,\mathbf{x}_2) & \cdots & \phi(\mathbf{x}_n,\mathbf{x}_n) \end{bmatrix}$$

where $\phi_{ij} = \phi(\mathbf{x}_i,\mathbf{x}_j) = \exp(-d(\mathbf{x}_i,\mathbf{x}_j)^2 / 2\sigma^2)$ and $d(.)$ is a distance metric. That is, the element ϕ_{ij} reflects the normalized distance between two data vectors of $(\mathbf{x}_i, \mathbf{x}_j)$ into the range [0, 1] by the Gaussian. Thus, the quantity of ϕ_{ij} becomes closer to 1 when $\mathbf{x}_i$ gets closer to $\mathbf{x}_j$ .Conversely, ϕ_{ij} becomes closer to 0 when the vector $\mathbf{x}_i$ gets farther from $\mathbf{x}_j$. Here the closeness between the two vectors is relatively defined by the width σ of the Gaussian. For a large σ, the Gaussian has a smooth shape incurring less impact of actual distance on the quantity of ϕ_{ij}. On the other hand, for a small σ, the Gaussian has a sharp shape incurring more impact of the distance on the quantity ϕ_{ij}.

2.2 Definition of RC

Assuming that the dataset $\mathbf{D}$ has k ($< n$) generated clusters, its corresponding *RC*, denoted by $RC(\tilde{\mathbf{D}}_k)$, is defined as follows.

Definition of $RC(\tilde{\mathbf{D}}_k)$: For a given dataset $\mathbf{D}$ consisting of n data vectors, $RC(\tilde{\mathbf{D}}_k)$ is defined by

$$RC(\tilde{\mathbf{D}}_k) = 1 - \frac{\| \mathbf{\Phi} - \tilde{\mathbf{\Phi}}_k \|_2}{\| \mathbf{\Phi} \|_2} \quad \text{where } \tilde{\mathbf{\Phi}}_k = \Sigma_{i=1}^{k} s_i \mathbf{u}_i \mathbf{v}_i^T \tag{1}$$

Here $\tilde{\mathbf{D}}_k$ represents the data of k generated clusters by clustering algorithm and $\mathbf{\Phi}$ is the distribution matrix of $\mathbf{D}$. For $\tilde{\mathbf{\Phi}}_k$, s_i denotes the i^{th} singular values of $\mathbf{\Phi}$, and $\mathbf{u}_i$ and $\mathbf{v}_i$ denote the i^{th} left and right singular vectors, respectively.

2.3 Properties of RC

The definition of $RC(\tilde{\mathbf{D}}_k)$ in Formula (1) can be also described as,

$$RC(\tilde{\mathbf{D}}_k) = 1 - \frac{s_{k+1}}{s_1} \tag{2}$$

where s_1, s_{k+1} are the 1st and the $(k+1)^{\text{th}}$ singular values, respectively, of $\mathbf{\Phi}$.

Proof. Based on Theorem 2.3.1 and Theorem 2.5.3 (see [18] for reference), the 2-norm of $\mathbf{\Phi}$ is the square root of the largest eigen value of $\mathbf{\Phi}^T\mathbf{\Phi}$, which is equal to the first singular value of $\mathbf{\Phi}$. Thus, $\left\|\mathbf{\Phi}\right\|_2 = s_1$. Also, $\left\|\mathbf{\Phi} - \tilde{\mathbf{\Phi}}_k\right\|_2 = s_{k+1}$, which completes the proof. □

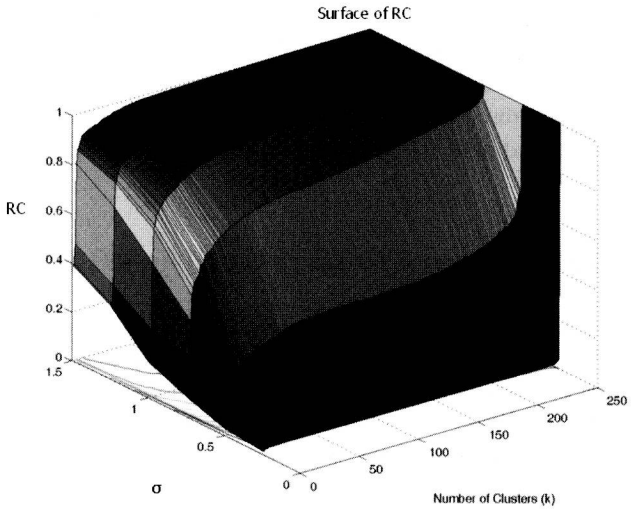


Fig. 1(a). The surface of *RC* simulated with the synthetic data for different choices of number of clusters *k*=(5:5:250) and the closeness parameter σ=(0.25:0.25:1.5)

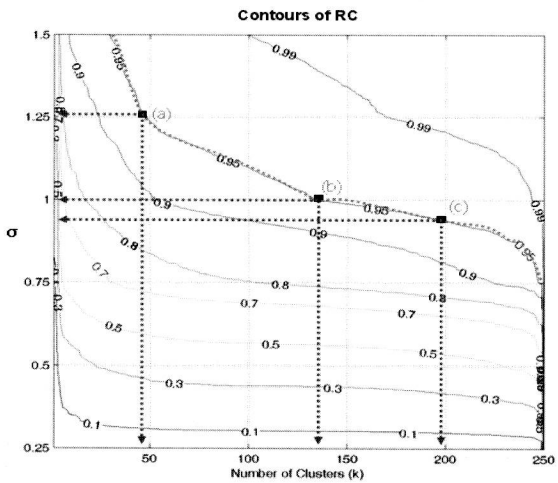


Fig. 1(b). The contours of RC simulated with the synthetic data for different choices of number of clusters k=(5:5:250) and the closeness parameter σ=(0.25:0.25:1.5)

Figures 1 (a) and (b) shows the surface and contours of *RC*, respectively, which have been simulated with synthetic data over the (k,σ) space. Here k is the number of clusters and σ is the closeness parameter. As seen in Figure 1(a), a larger number k of clusters increases the corresponding *RC* continuously up to reach a certain number of clusters and then it stays almost flat even with additional number of clusters, although the saturation point is dependent on the choice of σ. Also, in Figure 1(b), it is seen that to meet a certain level of the *RC*, a larger σ requires smaller number of clusters while a smaller σ requires larger number of clusters. Therefore, it is observed that for different choices of σ, we can have several different k's satisfying a specific RC criterion. For example, three possible choices of (k,σ) combinations denoted as (a), (b), and (c) are illustrated in the contours of Figure 1(b), where all of them have *RC* = 0.95.

3 Estimation of Number of Clusters Based on RC Criterion

In this section, we address the problem of estimating the good number of clusters for a given dataset. By using *RC*, it can be formulated as the problem of identifying the minimum number of clusters k such that its corresponding *RC* is not less than the specified *RC* having a certain amount of error allowance, say δ. Assuming an error allowance δ $(0 < \delta < 1)$, it means that we want to find "k" clusters having no less than $RC = 1-\delta$. Thus, in this case, the desired number of target clusters for a given dataset $\mathbf{D}$ should be the smallest k which satisfies the following condition:

$$RC(\tilde{\mathbf{D}}_k) \geq 1-\delta$$

By Formula (2), this can be also stated as

$$1-\frac{s_{k+1}}{s_1} \geq 1-\delta$$

That is,

$$\frac{s_{k+1}}{s_1} \leq \delta$$

$$s_{k+1} \leq s_1 \times \delta$$

$$s_k > s_1 \times \delta. \tag{3}$$

Now our concern is to find the smallest k satisfying condition (3). Interestingly, this problem corresponds to the problem of computing the *effective rank* (also called *ε-rank*) of the distribution matrix $\mathbf{\Phi}$. Note that for some small $\varepsilon > 0$, the *ε-rank* of a matrix $\tilde{\mathbf{D}}$ is defined as

$$r_\varepsilon = rank(\tilde{\mathbf{D}}, \varepsilon) \tag{4}$$

such that

$$s_1 \geq \cdots \geq s_{r_\varepsilon} > \varepsilon > s_{r_\varepsilon+1} \geq \cdots \geq s_n .$$

Here ε denotes the effect of noise and rounding errors in the data. Such a rank estimate r_ε is referred to as the *effective rank* of the matrix [18]. Putting the condition (3) into the *effective rank* definition shown in (4), therefore, the desirable number of clusters (k) can be obtained by estimating the ε-*rank* of $\mathbf{\Phi}$ with taking $\varepsilon = s_1 \times \delta$.

As a consequence, for the distribution matrix $\mathbf{\Phi}$ of a given dataset $\mathbf{D}$, the minimum number k of clusters satisfying a given condition of $RC(\tilde{\mathbf{D}}_k) \geq 1-\delta$ can be computed by

$$k = rank(\mathbf{\Phi}, \varepsilon) = rank(\mathbf{\Phi}, s_1 \times \delta) .$$

4 Experimental Results

Our experiments have been performed with two datasets, a synthetic dataset and a yeast cell-cycle dataset, for both of which the true numbers of clusters are already known along with the target clusters. For our analyses, the value of a closeness parameter σ has been heuristically chosen as in the range of $0< \sigma < \sqrt{d}/2$, where d is the dimensionality of data vectors. With different choices of error allowance δ =0.01, 0.05, 0.1, 0.2, and 0.3, we first identified the minimum number of clusters k to satisfy a given *RC* criterion, i.e. 1-δ , for the values of σ in the given range, and then generated the clusters with such a chosen k. The clustering results were then evaluated with the *adjusted rand index* as a validation index. Euclidean distance was used as a distance metric.

4.1 Experiment Methodology

4.1.1 Dataset

Synthetic data: The synthetic dataset was generated based on five different predefined time series patterns, which were partially taken from the nine synthetic time series patterns studied in [19]. This dataset includes 250 gene expression profiles consisting of their log expression measures at 10 different time points. For each of the five predefined patterns, 50 data vectors were uniformly generated by adding Gaussian noise $N(0,0.5^2)$ to it.

***Yeast cell cycle data*:** The real dataset used for our experiments is regarding mRNA transcript levels during the cell cycle of the budding yeast *S. cerevisiae.* In [20], Cho *et al.* monitored the expression levels of 6220 genes over two cell cycles, which were collected at 17 time points taken at 10 min intervals. Out of these genes, they identified 416 genes showing the peak at different time points and categorized them into five phases of cell cycle, viz. early G1, late G1, S, G2, and M phases. Among these,

by removing such genes that show the peak at more than one phase of cell cycle, 380 genes were identified and used in our experiments, whose expression levels clearly show the peak at one of the five phases of cell cycle.

4.1.2 Cluster Generation

For cluster generation, we used the seed-based clustering method which has been recently developed in [21]. The seed-based clustering method consists of two phases: seed extraction and cluster generation. The first phase of seed extraction is, given the number k of clusters, to find k good seeds of data vectors by computational analysis of given data matrix in such a way that the chosen seeds can be distinguished enough not to be very similar to each other while capturing all the unique data features (see [21] for more details). Once the seeds are chosen, the second phase proceeds to generate the clusters by using the chosen seeds as the representative vectors of potential clusters and assigning each data vector to a cluster with the closest representative vector. That is, by assigning each of the data vectors included in the dataset to the cluster of which representative vector is the most similar to the current data vector, the cluster memberships of all the data vectors are identified.

4.1.3 Cluster Assessment

Here clustering results are assessed by *adjusted rand index* (hereafter ARI), which is a statistical measure to assess the agreement between two different partitions and has been used in some previous research on gene expression data analysis [22]. The adjusted rand index is defined as in Formula (5), where a value closer to 1 implies that the two partitions are closer to perfect agreement.

Suppose that $U = \{u_1, \ldots, u_R\}$ is the true partition and $V = \{v_1, \ldots, v_C\}$ is a clustering result. Then, according to [6], the adjusted rand index is defined as follows:

$$\frac{\sum_{i,j}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right] - \left[\sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2}\right] / \binom{n}{2}} \tag{5}$$

where n is the total number of genes in the dataset, n_{ij} is the number of genes that are in both class u_i and cluster v_j, and $n_{i.}$ and n_j are the number of genes in class u_i and cluster v_j, respectively.

4.2 Analysis Results on Synthetic Data

For the synthetic data, we chose the closeness parameter σ in the range of σ=(0.25:0.25:1.5). Recall that the value of σ is heuristically determined in the range of $0 < \sigma < \sqrt{d}/2$, where d is the dimensionality of data vectors. Since the number of conditions in the synthetic data is 10, the range of σ was chosen as $0 < \sigma < \sqrt{10}/2$, that is 1.581. Table 1 shows numerically the *RC*-based automatically chosen number of