BIOTECHNO 2017

The Ninth International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies

May 21 - 25, 2017

Barcelona, Spain

BIOTECHNO 2017 Editors

Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, Germany

BIOTECHNO 2017

Foreword

The Ninth International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies (BIOTECHNO 2017), held between May 21 - 25, 2017 - Barcelona, Spain, covered these three main areas: bioinformatics, biomedical technologies, and biocomputing.

Bioinformatics deals with the system-level study of complex interactions in biosystems providing a quantitative systemic approach to understand them and appropriate tool support and concepts to model them. Understanding and modeling biosystems requires simulation of biological behaviors and functions. Bioinformatics itself constitutes a vast area of research and specialization, as many classical domains such as databases, modeling, and regular expressions are used to represent, store, retrieve and process a huge volume of knowledge. There are challenging aspects concerning biocomputation technologies, bioinformatics mechanisms dealing with chemoinformatics, bioimaging, and neuroinformatics.

Biotechnology is defined as the industrial use of living organisms or biological techniques developed through basic research. Bio-oriented technologies became very popular in various research topics and industrial market segments. Current human mechanisms seem to offer significant ways for improving theories, algorithms, technologies, products and systems. The focus is driven by fundamentals in approaching and applying biotechnologies in terms of engineering methods, special electronics, and special materials and systems. Borrowing simplicity and performance from the real life, biodevices cover a large spectrum of areas, from sensors, chips, and biometry to computing. One of the chief domains is represented by the biomedical biotechnologies, from instrumentation to monitoring, from simple sensors to integrated systems, including image processing and visualization systems. As the state-of-the-art in all the domains enumerated in the conference topics evolve with high velocity, new biotechnologes and biosystems become available. Their rapid integration in the real life becomes a challenge.

Brain-computing, biocomputing, and computation biology and microbiology represent advanced methodologies and mechanisms in approaching and understanding the challenging behavior of life mechanisms. Using bio-ontologies, biosemantics and special processing concepts, progress was achieved in dealing with genomics, biopharmaceutical and molecular intelligence, in the biology and microbiology domains. The area brings a rich spectrum of informatics paradigms, such as epidemic models, pattern classification, graph theory, or stochastic models, to support special biocomputing applications in biomedical, genetics, molecular and cellular biology and microbiology. While progress is achieved with a high speed, challenges must be overcome for large-scale bio-subsystems, special genomics cases, bio-nanotechnologies, drugs, or microbial propagation and immunity.

We take here the opportunity to warmly thank all the members of the BIOTECHNO 2017 Technical Program Committee, as well as the numerous reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to BIOTECHNO 2017.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the BIOTECHNO 2017 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that BIOTECHNO 2017 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the fields of bioinformatics, biocomputational systems and biotechnologies.

We also hope that Barcelona provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city.

BIOTECHNO 2017 Chairs:

Gilles Bernot, University Nice Sophia Antipolis, France
Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, Germany
Hesham H. Ali, University of Nebraska at Omaha, USA
Erliang Zeng, University of South Dakota, USA
Y-h. Taguchi, Chuo University, Japan
Hunter Moseley, University of Kentucky, USA
Magnus Bordewich, Durham University, UK

BIOTECHNO Industry/Research Advisory Committee
Steffen Heber, North Carolina State University, USA
Alexandru Floares, SAIA Institute, Romania

BIOTECHNO 2017

Committee

BIOTECHNO Steering Committee

Gilles Bernot, University Nice Sophia Antipolis, France
Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, Germany
Hesham H. Ali, University of Nebraska at Omaha, USA
Erliang Zeng, University of South Dakota, USA
Y-h. Taguchi, Chuo University, Japan
Hunter Moseley, University of Kentucky, USA
Magnus Bordewich, Durham University, UK

BIOTECHNO Industry/Research Advisory Committee

Steffen Heber, North Carolina State University, USA
Alexandru Floares, SAIA Institute, Romania

BIOTECHNO 2017 Technical Program Committee

Hesham H. Ali, University of Nebraska at Omaha, USA
Andrés Becerra Sandoval, Independent researcher, Colombia
Vincenzo Belcastro, Philip Morris International, Neuchatel, Switzerland
Gilles Bernot, University Nice Sophia Antipolis, France
Razvan Bocu, Transilvania University of Brasov, Romania
Magnus Bordewich, Durham University, UK
Peter Clote, Boston College, USA
Alexandru Floares, SAIA Institute, Romania
Xin Gao, King Abdullah University of Science and Technology (KAUST), Saudi Arabia
Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, Germany
Radu Grosu, Vienna University of Technology, Austria
Steffen Heber, North Carolina State University, USA
Asier Ibeas, Universitat Autònoma de Barcelona, Spain
Sohei Ito, National Fisheries University, Japan
Xiaoqian Jiang, UC San Diego, USA
Attila Kertesz-Farkas, National Research University - Higher School of Economics (HSE), Moscow, Russia
Cedric Lhoussaine, lab CRIStAL | University Lille 1, France
Chen Li, Monash University - Melbourne, Australia
Yiheng Liang, University of Central Missouri, USA
Giancarlo Mauri, University of Milano-Bicocca, Italy
Chilukuri K. Mohan, Syracuse University, USA
José Manuel Molina López, Universidad Carlos III de Madrid, Spain
Hunter Moseley, University of Kentucky, USA
Leif Peterson, Houston Methodist Research Institute, USA
Bhanu Rekepalli, BioTeam, USA
Vincent Rodin, University of Brest (UBO), France
J. Cristian Salgado Herrera, University of Chile, Chile

Thomas Schmid, Universität Leipzig, Germany

Jaime Seguel, University of Puerto Rico at Mayaguez, Puerto Rico

Patrick Siarry, Université Paris-Est Créteil, France

Christine Sinoquet, University of Nantes, France

Y-h. Taguchi, Chuo University, Japan

Sophia Tsoka, King's College London, UK

Marcel Turcotte, School of Electrical Engineering and Computer Science (EECS), Ottawa, Canada

Yanshan Wang, Mayo Clinic, Rochester, USA

Bin Xue, University of South Florida, USA

Erliang Zeng, University of South Dakota, USA

Qiang Zhu, The University of Michigan, Dearborn, USA

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Cancer Gene Analysis using Small Matryoshka (SM) Found by Matryoshka Feature Selection Method

Shuichi Shinmura

Faculty of Economics at Seikei Univ.
Tokyo, Japan
e-mail: shinmura@econ.seikei.ac.jp

Abstract— **We established a new theory of discriminant analysis after R. Fisher. We developed two methods and four Optimal Linear Discriminant Functions (OLDFs) in order to solve five serious problems of discriminant analysis. Over than 30 years, many researchers could not select cancer genes from gene datasets such as microarray datasets (Problem 5). The Matryoshka feature selection method (Method 2) could separate the dataset to several linearly separable subspaces (Small Matryoshka, SM) and non-linearly separable subspace definitely. We consider genes including each SM are cancer genes because they can discriminate cancer patients versus normal patients completely. On the other hand, other genes cannot discriminate two classes correctly and are noise. Therefore, Method 2 can separate the dataset to several signal SMs and noise subspace naturally. In this study, we introduce how to analyze all SMs of six datasets using common statistical methods and propose malignancy indexes for cancer gene diagnosis. Because our standard statistical approach obtains almost the same successful results, we explain the results by Alon et al. dataset. Researchers can analyze their dataset by our approach.**

Keywords—Cancer Gene Analysis; Cancer Gene Diagnosis; Microarray Dataset; Matryoshka Feature Selection Method; Small Matryoshka (SM); Basic Gene Set (BGS); NP-hard; large p small n.

I. INTRODUCTION

We established a new theory, which in this paper will be referred to as Theory, of discriminant analysis [28] after Fisher [8] and solved five serious problems of discriminant analysis [16][19]. We developed four Optimal Linear Discriminant Functions (OLDFs) and two new methods. Integer Programming (IP) defines IP-OLDF and Revised IP-OLDF (RIP) based on Minimum NM (MNM) instead of Number of Misclassifications (NM). Linear Programming (LP) defines Revised LP-OLDF. Revised IPLP-OLDF is a mixture model of Revised LP-OLDF and RIP. IP-OLDF found two new facts about discriminant analysis such as 1) the relation of NM and LDF, 2) MNM monotonic decrease ($MNM_k >= MNM_{(k+1)}$).

In Section 2, we explain the Matryoshka feature selection method achieved by LINGO [14] Program 3, which we call Method 2. Program 3 found all SMs of six datasets in Table 1. In Section 3, we analyze all 64 SMs of Alon et al. dataset [1] by common statistical methods. Those are one-way ANOVA with t-test, Ward cluster analysis, Principal Component Analysis (PCA), logistic regression [5][7], Fisher's Linear Discriminant Function (LDF) and a Quadratic Discriminant Functions (QDF). In Section 4, we analyze 64 RIP discriminant score data of 64 SMs; we get straightforward and surprising results. In Section 5, we discuss the reason why statistical methods cannot analyze the datasets with noise because of the large p small n problem [6] and NP-hard [3]. In Section 6, we conclude that RIP discriminant scores data is very useful because we can propose malignancy indexes for cancer gene diagnosis.

II. NEW THEORY OF DISCRIMINANT ANALYSIS

In this section, we introduce Theory outlook and all SMs of six datasets found by LINGO Program 3 [28].

A. Five Serious Problems of Discriminant Analysis

The five problems that we address in this paper are the following. Only RIP can discriminate the cases on the discriminant hyperplane correctly (we will refer to this as Problem 1). Because other LDFs cannot discriminate those cases correctly, their NMs may increase. Other LDFs, except for RIP and a Hard margin SVM (H-SVM) [35], cannot discriminate linearly separable data (LSD) (we will refer to this as Problem 2). Moreover, error rates of Fisher's LDF are very high for LSD discrimination. Problem 3 is the defect of the generalized inverse. QDF misclassifies all cases in class 1 to class 2 if some variable including class 1 is constant and its values including class 2 vary. If we add slight random numbers to the variable, we can solve Problem 3. The discriminant analysis is not traditional inference statistics that offer standard error (SE) equation derived from the normal distribution (we will refer to this as Problem 4). The 100-fold cross-validation method (Method 1) offers the 95% confidence intervals (CIs) of error rates and discriminant coefficients by the computer-intensive approach [15]. Over more than 30 years [10], many medical and statistical researchers were struggling to select cancer genes from the high-dimensional datasets with noise (we will refer to this as Problem 5). We call all linearly separable gene space as Matryoshka. We downloaded six microarray datasets from Higgins HP [12] on October 28th, 2015. When we discriminate these datasets, all NMs of three Revised OLDFs are zero and few coefficients less than case number "n" are not zero. We will refer to this subspaces as first Small Matryoshka (SM1). Next, when we discriminate

the modified dataset with SM1 removed, we find second SM (SM2). If we cannot find other SM anymore, we stop this iteration. This process is Method 2. Therefore, Program 3 finds the microarray dataset is disjoint unions of several SMs and non-linearly separable subspace. Because six microarray datasets are LSD, we believed other microarray datasets are LSD and have the Matryoshka structure, also. Moreover, "MNM monotonic decrease" explains the Matryoshka structure of LSD. Until now, there are no clear explanations about cancer genes. Most researchers do not know the microarray dataset is LSD. We consider genes including each SM as "cancer genes" because MNM using these genes is zero and we can discriminate cancer patients versus normal patients. Therefore, Program 3 can separate the microarray dataset into multiple signals and noise by only discriminating the dataset. Researchers at LASSO [31] [34] tried to zero some discrimination coefficients, and there were many filtering methods [2], but Method 2 has both abilities of LASSO and filtering method. Statistical methods could not be successful by analyzing noise containing microarray datasets, but these SMs are small samples and can be analyzed very easy. In this study, we introduce how to analyze all SMs by the standard statistical approach, and propose several malignancy indexes for cancer gene diagnosis.

B. Six MP-based LDFs and Two Statistical LDFs

Although we developed a diagnostic logic of Electrocardiogram data by Fisher's LDF and QDF, our research was inferior to the decision tree logic developed by the medical doctor. This experience is our motivation to develop Theory. After many experiences of the discriminant analysis, we developed IP-OLDF expressed in (1). Because we fix the intercept of IP-OLDF to 1, IP-OLDF is defined in the p-dimensional coefficient space omitting the intercept. Although $y_i*({}^t\mathbf{x}_i\mathbf{b}+1)$ is discriminant scores, $y_i*({}^t\mathbf{x}_i\mathbf{b}+1) = 0$ is a linear hyperplane and divides discriminant space to two half planes such as plus half plane $(y_i*({}^t\mathbf{x}_i\mathbf{b}+1) > 0)$ and minus half plane $(y_i*({}^t\mathbf{x}_i\mathbf{b}+1) < 0)$. If we choose $\mathbf{b}_k$ in plus hyperplane as LDF, LDF such as $y_i*({}^t\mathbf{b}_k\mathbf{x}_i+1)$ discriminates $\mathbf{x}_i$ correctly because of $y_i*({}^t\mathbf{b}_k\mathbf{x}_i+1) = y_i*({}^t\mathbf{x}_i\mathbf{b}_k+1) > 0$. On the other hand, if we choose $\mathbf{b}_k$ in minus hyperplane, LDF misclassifies $\mathbf{x}_i$ because of $y_i*({}^t\mathbf{b}_k\mathbf{x}_i+1) = y_i*({}^t\mathbf{x}_i\mathbf{b}_k+1) < 0$.

$$\text{MIN} = \Sigma\, e_i; \quad y_i*({}^t\mathbf{x}_i\mathbf{b} + 1) >= -\, e_i; \qquad (1)$$

e_i: 0/1 integer variable corresponding to classified/misclassified cases.
y_i: 1/-1 for class1/class2 or object variable.
$\mathbf{x}_i$: p-independent variables.
$\mathbf{b}$: discriminant coefficients.

Because IP-OLDF has a defect if data is not a general position, we developed RIP that looks for the interior point of true Optimal Convex Polyhedron (OCP) defined in (2) directly, NM of which is MNM. Because b_0 is free variable, RIP is defined in (p+1)-dimensional coefficient space. If it discriminates $\mathbf{x}_i$ correctly, $e_i = 0$ and $y_i*({}^t\mathbf{x}_i\mathbf{b}+b_0) >= 1$. If it cannot discriminate $\mathbf{x}_i$ correctly, $e_i = 1$ and $y_i*({}^t\mathbf{x}_i\mathbf{b}+b_0) >= -$

9999. Although Support Vector (SV) for classified cases are $y_i*({}^t\mathbf{x}_i\mathbf{b}+b_0) = 1$, SV for misclassified cases are $y_i*({}^t\mathbf{x}_i\mathbf{b}+b_0) = -9999$. Therefore, we expect a discriminant score of misclassified cases to be less than -1; there are no cases within SV. Because there are no cases on the discriminant hyperplane, we must understand that the optimal solution is an interior point of OCP defined by IP-OLDF [15], NM of which is MNM. Because all LDFs except for RIP cannot solve Problem 1 theoretically, these LDFs must check the number of cases (h) on the discriminant hyperplane. Correct NM may increase (NM + h). If e_i is non-negative real variable, equation (2) changes Revised LP-OLDF. Revised IPLP-OLDF is a mixture model of Revised LP-OLDF in the first phase and RIP in the second phase.

$$\text{MIN} = \Sigma e_i; \quad y_i*({}^t\mathbf{x}_i\mathbf{b} + b_0) >= 1 - M* e_i; \qquad (2)$$

b_0: free decision variables.
M: 10,000 (Big M constant).

When we discriminate Swiss banknote data with six variables [29], IP-OLDF finds that two variables models, such as (X4, X6), are linearly separable. By the monotonic decrease of MNN, 16 MNMs including these two variables are zero among 63 models (= $2^6-1 = 63$). Other 47 MNMs are greater than 1. This fact is important for gene analysis because (X4, X6) can list all Matryoshkas. Therefore, (X4, X6) is called cancer Basic Gene Set (BGS).

C. Theory Outlook

We established Theory and solved five serious problems of discriminant analysis. Let us consider two-class discrimination with n-cases and p-variables. IP-OLDF is defined on p-discriminant coefficient space. N-linear hyperplanes made by n-cases values as coefficients divide this space into finite Convex Polyhedron (CP). All LDFs corresponding to interior points of CP have unique NM and misclassify the same NM cases. Therefore, the relation between NM and LDF coefficient was first clarified. If we consider that all LDFs corresponding to the same interior points are equivalent, there are finite LDFs as same as finite CPs. There is an OCP. MNM decreases monotonously $(MNM_k >= MNM_{(k+1)})$ because k-coefficient space is sub-space of (k+1)-coefficient spaces. If $MNM_k = 0$ and k is a minimum number of variables, all MNMs including these k-variable are zero. This fact means LSD has Matryoshka structure. Therefore, MNM is a critical statistic of LSD discrimination. We call the k-variable model a cancer BGS, that is the smallest SM in gene analysis because it can explain the Matryoshka structure of high-dimensional dataset by BGS. IP-OLDF finds the right vertex of OCP if data is a general position. However, it may not find the right vertex of OCP if data is not general position. Therefore, we develop RIP that looks for interior point of OCP directly. All LDFs except for RIP cannot discriminate the cases on discriminant hyperplane correctly if they choose the vertex or edge of CP. Therefore, NM of these LDFs may not be correct (Problem 1). Only H-SVM and RIP can discriminate LSD correctly (Problem 2). Although Revised LP-OLDF tends to collect the cases on the discriminant hyperplane, it can discriminate LSD correctly. Logistic regression almost discriminate LSD.

However, Fisher's LDF and QDF often cannot discriminate LSD. QDF and Regularized Discriminant Analysis (RDA) [9] misclassify all cases to another class because of the defect of a generalized inverse (Problem 3). Fisher proposed Fisher's LDF based on Fisher's assumption and established the discriminant theory based on variance-covariance matrices. Although many data do not satisfy Fisher's assumption and there is no real test statistics for it, Fisher's LDF solves many applications. Because there is no equation of SE of LDFs, the discriminant analysis is not the inferential statistical method (Problem 4). Therefore, we developed Method 1 [17] and LINGO Program 2 that can offer the 95% CIs of discriminant coefficients and error rates [18]. Moreover, we developed useful and simple model selection method such as the best model with a minimum mean of error rate in the validation samples (M2) [29]. After examining all the models, we concluded that the best models of RIP were almost better than other seven LDFs such as two OLDFs, three SVMs, logistic regression and Fisher's LDF. About Problem 5, we discuss in section D.

D. Problem 5 and Matryoshka Feature Selection Method

First of all, we developed Program 3 for RIP. However, Program 3 can discriminate the dataset with Revised LP-OLDF, H-SVM, and two Soft-margin SVMs (S-SVMs). Two S-SVMs are SVM 4 (penalty c = 10000) and SVM 1 (penalty c = 1). Revised LP-OLDF and Revised IPLP-OLDF can find SM1 and stop the iteration. Although NMs of three SVMs are zero, most coefficients are not zero. Therefore, three SVMs are not helpful for cancer gene selection because we must survey all possible models [11] to find SM and it is NM-hard. Because NMs of Fisher's LDF by JMP are not zero, Fisher's LDF is not useful for gene analysis. Therefore, only three Revised OLDFs could find that the dataset consisted disjoint unions of several SMs and another noise subspace. To the best of our knowledge, there was no absolute definition of cancer gene, and the purpose of cancer gene analysis was not clear until now. Now, we can define cancer gene set that can discriminate cancer patients versus normal patients or just two different types of cancer. Moreover, because Program 3 separated several signal SMs and noise subspace, Program 3 is a good filtering system that removes noise subspace from the dataset. Program 3 could find all SMs of six datasets in Table 1 [26]. Moreover, we confirmed and validated Program 3 using Swiss banknote data and Japanese automobile data [27] also. Therefore, we claim cancer gene analysis is very easy and exciting theme. When we discriminated Shipp et al. dataset [30] on Oct. 28, 2015, RIP could select 32 genes among 7129 genes. We thought the discrimination having 7129 variables needed huge CPU time by NP-hard. However, Fisher's LDF [13] and six MP-based LDFs [14] can solve the datasets in less than 20 seconds because the datasets are LSD. Generally speaking, MP-based six LDFs are difficult by NM-hard. If datasets are LSD, these LDFs are free from NP-hard. However, most coefficients of three SVMs are not zero. Therefore, SVMs are not helpful for feature selection for gene analysis in addition to common data. Because Revised

LP-OLDF minimizes the summation of misclassified case distance from the discriminant hyperplane, which is the second objective function of S-SVM, it can discriminate LSD correctly because this standard is the same as MNM standard only for LSD. However, this standard tends to gather cases on discrimination planes (Problem 1). The SV distance maximization criterion solved by the Quadratic Programming (QP) seems to be preventing the SVM discriminant coefficient from becoming zero. Table 1 shows the summary of six datasets found until December 20th, 2015. Rows "Size" are the case number by the gene number. Rows "SM: Gene" are the number of SM: the total number of genes including in all SMs. Six papers [20] - [25] include full gene name including each SM. Rows "JMP12" are two by two tables of the discrimination by Fisher's LDF. Six NMs are 5, 3, 8, 3, 10 and 29 and error rates are very high. If BGS has k genes, the dataset with p variables includes many smaller Matryoshka from (p - 1) variables to k variables. Program 3 finds that the datasets are the disjoint union of SMs with h-variables (p > h >= k) and another high-dimension gene subspace with "MNM >= 1." Now, we must survey the BGSs from SM by manual operation. If Revised LINGO Program 3 can find all list of BGSs, we can understand the Matryoshka structure of the dataset by these BGSs completely. Because we can analyze each SM using common statistical methods, we expect to obtain new facts of gene analysis and hope many researchers try to analyze these SMs. By our breakthrough, the cancer gene analysis becomes an interesting theme.

TABLE I.   SUMMARY OF SIX MICROARRAY DATASETS

Data	Alone et al. [1]	Chiaretti et al. [4]
Size	62 *2000	128*12625
SM: Gene	**64** [22]:**1999**	269 [25]:5220
JMP12	20:2 / 3:37	94:1 / 2:31
Data	Golub et al. [10]	Shipp et al. [30]
Size	72*7129	77 *7130
SM: Gene	67 [21]:1203	214 [20]:3040
JMP12	20:5 / 3:44	17:2 / 1:51
Data	Singh et al. [32]	Tian et al. [33]
Size	102 *12626	173 *12625
SM: Gene	178 [23]:3984	159 [24]:7221
JMP12	46:4/6:46	16:20/9:128

III. ANALYSIS OF 64 SMs

In this section, we analyze 64 SMs by common statistical methods because all SMs are small samples.

A. 64 SMs of Alon et al. Dataset found by Method 2

We discriminate Alon et al. [1] dataset by LINGO Program 3. Table 2 tells us the dataset consists of disjoint unions of 64 SMs from SM=1 to SM=64 and one gene. Therefore, Alon et al. were successful in separating signals and noise although they may drop some signals. "Gene"

column is the number of genes in each subspace. All NMs of logistic regression and QDF are zero; we omit two columns from the table. "LDF2 and LDF1" are NMs of two different prior probability options of Fisher's LDFs. The prior probabilities of LDF2 are proportional to the 22 cases in class 1 and 40 cases in class 2. Those of LDF1 are the same and are the default setting in most statistical software. However, we use the former prior probability because we wish to compare NMs of six MP-based LDFs. Ten NMs of LDF2 are greater than NMs of LDF1, and 13 NMs of LDF2 are less than NMs of LDF1. "SR" is the range of RIP discriminant scores. "RatioSV" is the value calculated by 2/SR*100 % that indicates the ratio of the distance of SV for the range of RIP discriminant scores. "t" column is t-value of the discriminant score. Last four rows are maximum, mean, minimum and summation of 64 SMs. The range of Gene is [21, 42], and 64 SMs include 1999 genes. The ranges of LDF2 and LDF1 are [0, 8] and [0, 9], respectively. The 13 NMs of LDF2 are zero, and 12 NMs of LDF1 are zero. The ranges of SR and RatioSV are [7.5, 84.9] and [2.4%, 26.8%], respectively. The maximum RatioSV 26.8% means SV width is 26.8% of RIP discriminant scores and separates two classes. Moreover, the 23 values of RatioSVs are over 5%. We must survey the proper threshold for malignancy indexes by validation in the near future. In this paper, we assume the threshold is 5% or 10%. The range of t-values is [-1.1, 4.6]. Nevertheless, 64 SMs are linearly separable; t-test is not helpful to find linearly separable signs.

TABLE II.   SUMMARY OF 64 SMS (OMITTED 54 SMS)

SM	GENE	LDF2	LDF1	SR	RatioSV	t
8	31	0	0	7.5	**26.8**	**0.3**
1	29	2	2	13.5	14.8	2.3
2	33	1	1	14.7	13.6	1.7
60	34	5	6	18	11.1	4.2
4	27	3	3	20.8	9.6	1
44	33	3	4	21.8	9.2	2.5
45	29	1	2	21.7	9.2	1.6
37	30	5	5	23.8	8.4	1.9
14	26	0	0	39.8	**5**	0.5
64	42	8	9	84.9	**2.4**	3.5
Max	42	8	9	84.9	26.8	4.6
Mean	30.8	2.1	2.2	19	12.8	1.7
Min	21	0	0	7.5	2.4	-1.1
SUM	1999	134	139	1218.5	821.6	109.8

B. Histogram and Correlation

Because 64 SMs of logistic regression and QDF are zero, these discriminant functions confirm 64 SMs are linearly separable and the quality of Program 3.
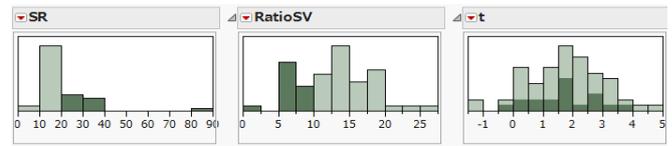
Fig. 1. The Histograms of Gene, LDF2, SR, RatioSV and t.

Figure 1 shows the histograms of SR, RatioSV, and t-value. We select 19 SMs with "RatioSV < 10% and show the portion of 19 SMs by the dark green. The "SR >= 20" are the same as 19 SMs with "RatioSV < 10%. These RIPs using 19 SMs probably need the validation by Method 1 because 19 RIPs may misclassify some patients. Dark green t-value almost cover the ranges. Although most researchers use the t-test, we think t-test is not helpful for gene analysis because we cannot find the useful meaning of all genes included in 64 SMs by t-tests.

C. Ward Cluster Analysis

Because four NMs of logistic regression, QDF, LDF2, and LDF1 are zero, we focus on this SM9. We analyze the dataset by Ward cluster analysis. Figure 2 is the heat map of SM9 with 33 genes. Right dendrogram shows cases, and lower dendrogram shows variables. It tells us Ward cluster cannot classify two classes by two clusters clearly. We categorize two clusters. We omit upper cluster including 33 cases that consist of 13 normal cases marked by □ and 20 cancer patients marked by ×. The lower cluster includes 29 cases that include nine normal and 20 cancer cases. This result shows the cluster analysis is not helpful for gene analysis.
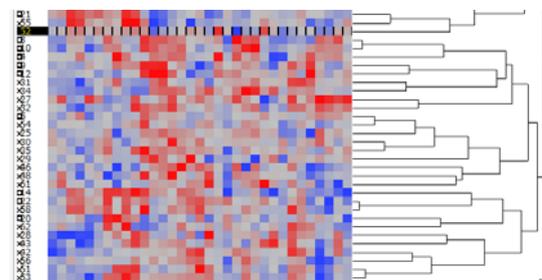
Fig. 2. Heat Map and Dendrogram of Cases.

Figure 3 is the dendrogram of 33 genes. We expect the specialist explains this result. Due to the small distance between X471 and X201, these genes may be able to replace other genes with each other.
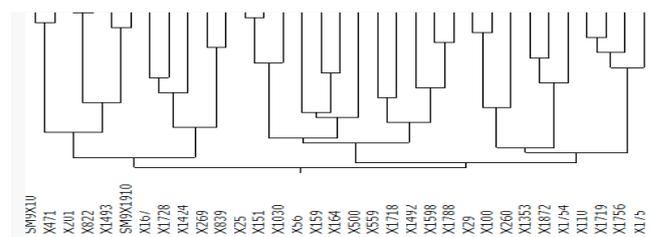
Fig. 3. Dendrogram of Genes.

Figure 4 is PCA output. Left figure is an eigenvalue. Nine principal component eigenvalues are greater than 1. The middle figure is a scatter plot. The right figure is factor loading plot. Although MNM of two classes is zero, scatter plot shows two classes overlap. We confirm 64 SMs scatter plots overlap and cannot find the linearly separable sign. We conclude it is difficult to find the useful meaning of these results. Comparison with Figure 6 confirms our claim.
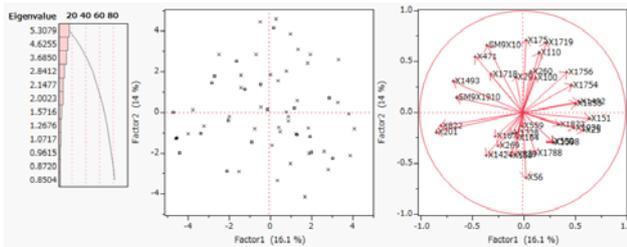
Fig. 4. Three Figures of PCA.

IV. ANALYSIS OF 64 RIP DISCRIMINANT SCORES

We claimed we could analyze SM by common statistical methods. However, we cannot obtain useful results. In this section, we analyze 64 RIP discriminant scores data by common statistical methods and obtain the surprising results. The data consists 62 patients (cases) with 64 discriminant scores (variables).

A. Examination of 64 RIP Discriminant Scores and RarioSV

Table 3 shows the summary of 64 RIP discriminant scores. Min and Max columns are the range of normal class 1. MIN and MAX are the range of tumor class 2. SR and RatioSV are the same in Table 2. The 22 normal cases are less than or equal to -1, and the 40 cancer cases are greater than or equal to 1. RIP SV separate two classes of 64 SMs completely. RatioSVs of SM8, SM14 and SM64 are 26.76%, 5.03% and 2.35%, respectively. We guess that the 63 RIPs whose RatioSVs are over 5% may be good malignancy indexes for cancer gene diagnosis.

TABLE III. SUMMARY OF 64 DISCRIMINANT SCORES OMIT 59 SM)

SM	Min	Max	MIN	MAX	SR	RatioSV
8	-3.35	-1	1	4.12	7.47	**26.76**
35	-2.58	-1	1	5.92	8.51	23.52
11	-4.15	-1	1	5.52	9.67	20.68
59	-15.25	-1	1	21.91	37.17	5.38
14	-21.94	-1	1	17.85	39.79	**5.03**
64	-3.94	-1	1	81	84.94	**2.35**
MAX	**-2.58**	-1	1	**81**	84.94	26.76
MIN	**-21.94**	-1	1	**4.12**	7.47	2.35

B. Ward Cluster Analysis and PCA

Ward cluster analyzes the discriminant scores data that consists of 62 patients (cases) with 64 RIP discriminant scores (variables); see Figure 5. The upper cluster is 22 normal cases, and the lower cluster is 40 cancer cases. Ward cluster separates two classes. However, it cannot separate two classes in Figure 2. The 62 cases dendrogram becomes over ten clusters. The 64 discriminant scores dendrogram has more complex clustering.
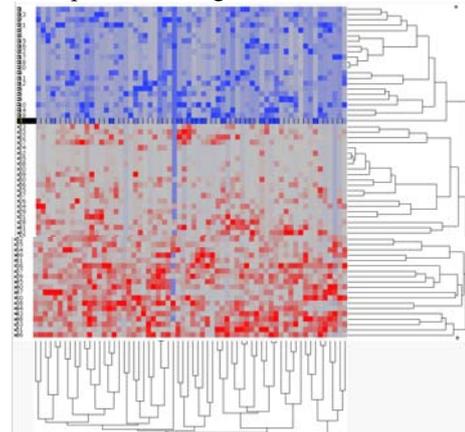
Fig. 5. Ward Cluster Analysis.

Figure 6 is three plots of PCA. The eigenvalue of the first principal component (Prin1) is enormous. Scatter plot shows two classes are completely separable. The 22 normal cases are on a negative axis of Prin1. The 40 cancer cases scatter on the first and fourth quadrants that look like a fan, same as factor loading plot. If we obtain the validation cases, we can judge whether the scatter plot is useful for malignancy index. Although we could not separate two classes in Figure 4, we can separate two classes in Figure 6 very clearly.
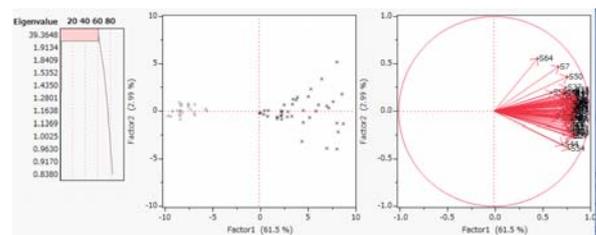
Fig. 6. Three Figures of PCA.

Table 4 is the ranking of four principal components. "R1" is the ranking of Prin1 of 62 cases in descending order. We expect "Prin1" to be the malignant degree of cancer. Because the Prin1 values of tumor class are greater than 0.02 and those of normal class are less than -5.55, RatioSV of Prin1 is 30.37% (= (0.02+5.55) / (8.77+9.57) * 100). We believe cancer patients of ID = 46, 36 and 53 are malignant patients. Moreover, normal patients of ID =18, 20, 17 may become cancer. If medical doctors confirm these patients are serious patients and normal patients with risk (possibility to

become cancer), we can use PCA as a cancer diagnosis. If these indexes misclassify cancer patients cured with some treatment to the normal class, the doctor can decide that the patient is cured ~~earlier than the diagnosis by~~ before 5-years ~~survival rate~~ because their measurements belong to the range of normal class.

TABLE IV.   MALIGNANCY INDEX OF PCA

ID	Prin1	R1	Prin2	R2	Prin3	R3	Prin4	R4
46	**8.77**	**1**	-1.35	58	-2.17	59	-1.29	56
36	**8.69**	**2**	1.8	4	2.62	5	0.47	18
53	**8.29**	**3**	-2.12	59	0.83	14	-0.91	47
58	0.43	38	-0.31	40	-0.14	36	-0.02	30
52	0.14	39	-0.19	37	0	30	0.43	20
55	**0.02**	40	-0.23	39	-0.2	37	0.29	25
18	**-5.55**	**41**	0.01	29	0.1	27	0.03	28
20	**-5.64**	**42**	0.22	22	0.13	25	-0.05	31
17	**-6.8**	**43**	-0.43	43	-0.81	47	-0.19	38
21	-9.12	60	-0.07	35	-0.05	32	0.72	14
10	-9.33	61	-0.62	51	0.15	23	-1.05	52
4	-9.57	62	-0.52	44	0.02	29	-1.04	51

C.   Analysis of Transpose Data

We examine three figures of PCA using transpose data with 64 discriminant scores (cases) by 62 cases (variables) in Figure 7. Factor loading plot shows 22 normal cases (variable) locate in the 2nd quadrant, and 40 cancer cases (variables) locate in the first and fourth quadrants. Scatter plot shows most discriminant scores are placed on the line of -45 degrees with the Prin1. The 13th, 14th, 28th, 63th and 64th discriminant scores are outliers. The first eigenvalue is not large compared with Figure 9.
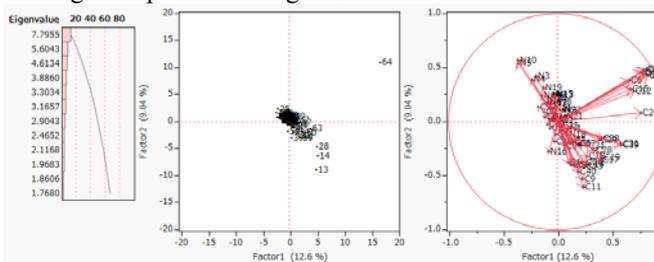
Fig. 7. Three Figures of PCA.

Figure 8 is four scatter plots of PCA. The x-axis is the Prin1. Y-axes are Prin2 and Prin3.
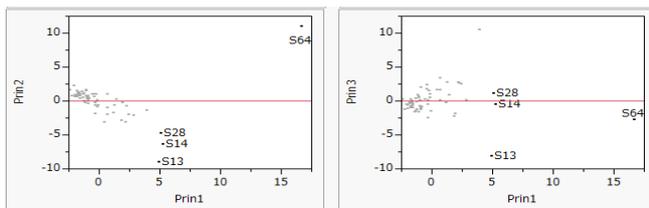
Figure 8. Four Figures of PCA.

There are four RIPs different from other RIPs. These RIPs may show different malignancy indexes. This remains to be validated by medical specialists in future work

D.   Summary of Six Datasets

Table 5 is the overview of six datasets. Alon et al. and Singh et al. are normal class versus cancer class discrimination. Other four datasets are just two different types of cancer discrimination. However, we obtained almost the same results as 64SMs of Alon et al. dataset. "SM" column shows the number of SMs. Because we find 130 BGSs of Alon et al. dataset, we analyze 130 BGSs in the first row. The range of 130 RIP RatioSVs and 64 RIP SMs are [0.00%, 0.9%] and [2.4%, 26.8%], respectively. Therefore, 130 RatioSVs of BGS are not helpful for malignancy indexes. However, we expect medical specialists examine the gene combination including each BGS and find useful results. ">=5%" column shows the number of SMs with RatioSV over 5%. All 95 RatioSVs of Chiaretti et al. are greater than 5%. Most RatioSVs of Alon et al. and Shipp et al. are greater than 5%, also. Therefore, we conclude Alon et al., Chiaretti et al. and Shipp et al. discriminate two classes easier than other three datasets. "PCA" column shows the RatioSVs of Prin1 that are greater than maximum RatioSVs of RIPs. All NMs of logistic regression are zero. "QDF and LDF2" columns are the number of NM=0 of QDF and LDF2. Although logistic regression and QDF discriminate 159 and 158 SMs by NM=0, Tian et al. dataset has 27 SMs with ">=5%". Although logistic regression and QDF discriminate 159 and 158 SMs by NM=0, Tian et al. [33] has 27 SMs with ">=5%". Tien et al. dataset is easier to discriminate two classes by logistic regression and QDF. Nevertheless, Fisher's LDF are not useful for the datasets except for Chiaretti et al.; statistical discriminant functions can sometimes discriminate SMs correctly.

TABLE V.   SUMMARY OF SIX MICROARRAY DATASETS

Data	SM	>=5%	PCA	QDF	LDF2
Alon et al. (BGS)	130	0	4.50%	60	0
Alon et al.	64	63	30.40%	64	13
Singh et al.	179	38	14.40%	26	0
Golub et al.	69	13	34.90%	16	1
Tien et al.	159	27	24.00%	158	1
Chiaretti et al.	**95**	95	51.50%	95	**92**
Shipp et al.	130	129	31.70%	121	46

Figure 9 is three plots of PCA using Chiaretti et al. dataset. Eigenvalue of Prin1 is gigantic compared with Fig. 7 because we guess two classes are apart from each other. Right factor loading plot locates on the first and fourth quadrants that look like a fan. Center scatter plot shows two

classes are completely separable. The 95 B-cell patients in class 1 locate on negative first principal axis (Prin1). The 33 T-cell patients locate on the positive Prin1. Figure 6 is a typical pattern of cancer prediction. On the other hand, Figure 9 may be a typical pattern of just two different types of cancer (cancer classification) introduced by Golub et al..
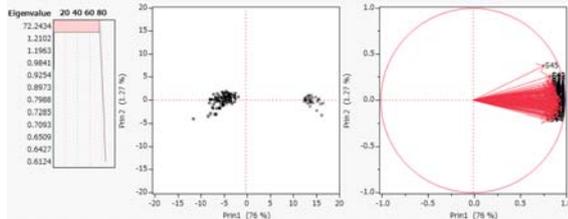
Fig. 9 Three plots of PCA (Chiaretti et al. dataset)

V. WHY IS CANCER GENE ANALYSIS DIFFICULT ?

Golub et al. said: "Although cancer classification has improved over the past 30 years, there has been no general approach for identifying cancer classes (class discovery) or for assigning tumors to known classes (class prediction)." We are unfamiliar with the history of cancer gene analysis and show "over ten years" in the book [28]. We knew some statisticians tried to discriminate gene data by Fisher's LDF about 20 years ago. Now, it is hard for us to find these studies because these trials were judged to be not useful for cancer gene analysis. We understood that most researchers are disappointed in the statistical discriminant analysis. At least, many scientists approached this theme by engineering methods [12] and restricted statistical methods, such as t-test, cluster analysis, and SVM. In this Section, we discuss the reason why cancer gene analysis was difficult.

Diao and Vidyashanker [6] explain cancer gene analysis "large p small n problem." In general, statistical methods treat small sample (small p small n problem) very easy. Now, statistical methods are difficult to analyze a big data (large p and large n) with noise. However, most statistical methods are easy to analyze the data with small p large n problem because of hardware and software ability improvement. Because one-way ANOVA with t-test analyzes each one variable, it is easy to analyze the datasets. On the other hand, some statistical methods are difficult to analyze the datasets with large p small n problem. Regression analysis and discriminant analysis based on variance-covariance matrices are difficult to the datasets with large p small n problem because it is hard to construct large p variance-covariance matrices using small n cases in addition to select feature. In Japan, although JMP released Fisher's LDF for large p small n problem in 2015, the error rate of six datasets are very high in Table 1 because Fisher's LDF cannot discriminate LSD correctly. Charikar et al. [3] introduced the problem called "combinatorial feature selection problems" is NP-hard. Their study gave a significant impact for gene feature selection researchers. In general, their claim is correct, especially for IP models such as RIP. However, six MP-based LDFs including RIP can discriminate each dataset in less than 20 seconds because the datasets are LSD. Until now, there was

no study about LSD discrimination. RIP found six datasets are LSD. LSD has the Matryoshka structure because it includes small Matryoshkas in it. Alon et al. and Singh et al. discriminate two classes such as cancer and normal, and other datasets discriminate just two different types of cancer. Both types of dataset consist disjoint unions of several SMs (signals) and other subspace (noise) that is not linearly separable. Our last goal is to find BGS in each SM. BGSs can explain the Matryoshka structure of each dataset uniquely. Medical gene specialist will be able to explain the useful meaning of the genes combination including all SMs shortly. However, we can show the useful results using RIP discriminant scores of 64 SMs of Alon et al.. We already obtain the same results of other five datasets in Table 5, also.

Why did many researchers spend to analyze the datasets over more than 30 years ? Our claims are as follows:

1) Many scientists analyze a dataset with noise by statistical methods. Because they could not obtain useful results, they trusted on engineering approaches, such as several filtering techniques. However, RIP can separate disjoint signals (SMs and BGSs) and noise subspace very easy.

2) If most gene data are LSD, the cancer gene analysis is very easy and straightforward. What is the cancer gene? The answer is very simple. It is the combination of genes including BGS. If we drop one gene from BGS, we can not discriminate this subspace correctly and misclassify some patients.

3) Because NMs of Fisher's LDF are not zero, it is not useful for cancer gene analysis. Because NMs of three SVMs are zero and most coefficients are not zero, SVM is not helpful for cancer gene analysis. If SVM computes all possible models, it can find BGS. However, this trial is NP-hard.

Our conclusion is as follows: Only RIP and Revised LP-OLDF can select gene feature naturally and find SMs. Now, we can find BGSs by manual operation after finding SMs. Every researcher can easily analyze SMs and BGs by standard statistical approaches. Moreover, we find several malignancy indexes. We must confirm our claim by the validation samples.

VI. CONCLUSION

We claimed common statistical methods could analyze SM easily because these subspaces were small samples [28]. However, our examination shows that it is difficult for us to obtain good results showed in Section 3. However, we can get clear results from the RIP discriminant scores in Section 4. Especially, the 63 RIPs with RatioSV over than 5% may be useful malignancy indexes for cancer gene diagnosis. If medical specialists examine and confirm our results and claims, our collaboration will open a new frontier of cancer gene diagnosis from cancer gene analysis. Moreover, the Ward cluster analysis can identify two clusters completely in

Figure 5. Usually, cluster analysis cannot cluster two classes clearly. Figure 6 and Table 3 of PCA show another reliable malignancy index. Healthy patients are located on the negative segment of the Prin1 axis, especially malignant tumor patients widely scatter in the first and fourth quadrants as in Figure 6. Furthermore, the scatter plot of Figure 8 shows the diversity of various tumor patients. We can find 64 SMs and analyze it successfully. However, these results are obtained using the statistical approach. We need co-operation with experts such as medical doctors as follows:

1) We already listed up all genes lists including SMs of six datasets. If the specialists check it, they may find the useful meaning of the combination of genes in each SM.

2) We expect the specialists examine the ranking on the Prin1 and confirm it shows the malignancy indexes.

3) If the specialists offer three types of validation samples such as normal cases, cancer patients and patients with cancer cured, we can discriminate those cases and expect the following result; 1) The normal cases are classified to the normal class. 2) The cancer patients are classified to the cancer class. 3) The patients with cancer cured are classified to the normal class. The misclassified number of the third group shows the degree of cure. If 63 RIPs misclassify the patients in the third group, they may be cured completely, and are relieved from the anxiety of five years.

REFERENCES

[1] U. Alon, et al., "Broad Patterns of Gene Expression Revealed by Clustering Analysis of cancer and Normal Colon Tissues Probed by Oligonucleotide Arrays," Proc. Natl. Acad. Sci. USA, 96, pp. 6745-6750. 1999.

[2] A. B. Brahim and M. Lima, "Hybrid Instance Based Feature Selection Algorithms for Cancer Diagnosis," Pattern Recognition Letters, pp.8. 2014.

[3] M. Charikar, et al., "Combinatorial feature selection problems," IEEE Xplore, pp. 631640. 2000.

[4] S. Chiaretti et al., "Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival," Blood. April 1, 2004, 103/7, pp. 2771-2778, 2004.

[5] D. R. Cox, "The regression analysis of binary sequences (with discussion)." J Roy Stat Soc B 20: pp. 215-242. 1958.

[6] G. Diao, and A. N. Vidyashankar. "Assessing Genome-Wide Statistical Significance for Large p Small n Problems," Genetics, 194, pp. 781–783, 2013.

[7] D. Firth, "Bias reduction of maximum likelihood estimates," Biometrika, vol. 80, pp. 27-39, 1993.

[8] R. A. Fisher, Statistical methods and statistical inference. Hafner Publishing Co. 1956.

[9] J. H. Friedman, "Regularized Discriminant Analysis," Journal of the American Statistical Association, 84/405, pp. 165-175, 1989.

[10] T. R. Golub et al., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," Science. 1999 Oct 15, 286/5439, pp. 531-537, 1999.

[11] J. H. Goodnight, SAS technical report – the sweep operator: its importance in statistical computing – R(100). SAS Institute Inc. 1978.

[12] I. B. Jeffery, D. G. Higgins, and C. Culhane, "Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data," BMC Bioinformatics. Jul 26 7:359, pp.1-16, Jul. 2006. doi: 10.1186/1471-2105-7-359.

[13] J. P. Sall, L. Creighton, and A. Lehman, JMP Start Statistics, Third Edition. SAS Institute Inc. 2004.

[14] L. Schrage, Optimization Modeling with LINGO. LINDO Systems Inc. 2006.

[15] S. Shinmura, The optimal linearly discriminant function. Union of Japanese Scientist and Engineer Publishing, Japan. 2010 (ISBN 978-4-8171-9364-3)

[16] S. Shinmura, "End of Discriminant Function based on Variance-Covariance Matrices," ICORES, pp. 5-14, 2014.

[17] S. Shinmura, "Comparison of Linear Discriminant Function by K-fold Cross-validation," Data Analytic 2014, pp. 1-6, 2014.

[18] S. Shinmura, "The 95% confidence intervals of error rates and discriminant coefficients," Statistics Optimization and Information Computing, 3, pp. 66-78, 2015.

[19] S. Shinmura, "Four Serious Problems and New Facts of the Discriminant Analysis," In E. Pinson, F. Valente, B. Vitoriano, (Eds.), Operations Research and Enterprise Systems, pp. 15-30, 2015. Springer (ISSN: 1865-0929, ISBN: 978-3-319-17508-9, DOI: 10.1007/978-3-319-17509-6).

[20] S. Shinmura, "Complete Lists of Small Matryoshka in Shipp et al. Microarray Data (9)," Research Gate (9), pp. 1-81, 2015.

[21] S. Shinmura, "Sixty-nine Small Matryoshka in Golub et al. Microarray Data (10)," Research Gate (10), pp. 1-58, 2015.

[22] S. Shinmura, "Simple Structure of Alon et al. et al. Microarray Data (11)," Research Gate (11), pp. 1-34, 2015.

[23] S. Shinmura, "Feature Selection of Singh et al. Microarray Data (12)," Research Gate (12), pp. 1-89, 2015.

[24] S. Shinmura, "Final List of Small Matryoshka in Tian et al. Microarray Data," Research Gate (13), pp. 1-160, 2015.

[25] S. Shinmura, "Final List of Small Matryoshka in Chiaretti et al. Microarray Data," Research Gate (14), pp. 1-16, 2015.

[26] S. Shinmura S, "Matroska Feature Selection Method for Microarray Data," Biotechno 2016, pp.1-8 2016 (Best Paper Award)

[27] S. Shinmura, "Discriminant Analysis of the Linearly Separable Data," Journal of Statistical Science and Application, 2016.

[28] S. Shinmura, New Theory of Discriminant Analysis after R. Fisher. Springer, Dec. 2016.

[29] S. Shinmura, "The Best model of the Swiss banknote data," Statistics, Optimization and Informatics Computing, 3 Spring 2016, pp. 1-24, 2016.

[30] M. A. Shipp et al., "Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning," Nature Medicine 8, pp. 68-74, 2002.

[31] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso," J. Comput. Graph. Statist., 22. pp. 231-245, 2013.

[32] D. Singh et al., "Gene expression correlates of clinical prostate cancer behavior," Cancer Cell: March 2002, Vol.1, pp. 203-209, 2002.

[33] E. Tian et al., "The Role of the Wnt-Signaling Antagonist DKK1 in the Development of Osteolytic Lesions in Multiple Myeloma," The new England Journal of Medicine, Vol. 349, 26, pp. 2483-2494, 2003.

[34] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," J. R. Statist. Soc. B 58/1, pp. 267-288, 1996.

[35] V. Vapnik, The Nature of Statistical Learning Theory. Springer. 1999.

Overlapping Cell Segmentation with Shape Prior

Evan Yi-Wen Wang, Albert Kai-Sun Wong

Department of Electronic and Computer Engineering
Hong Kong University of Science and Technology
Hong Kong
Email: {ywangda, eealbert}@ust.hk

Abstract—**Cell segmentation is a critical task in fully automatic computer cytology diagnosis. Overlapping cells pose a major challenge to cell segmentation because of blurred edges and inhomogeneous cytoplasm. In this paper, we propose a novel energy functional and formulate the boundary identification problem as a problem to minimize this energy functional together with a shape prior. The terms of our energy functional are designed to exploit both the edge and regional features, and the shape prior is derived from low rank optimization. For the non-convex minimization problem, we formulate an efficient monotone Accelerated Proximal Gradient with Linesearch (APGL) algorithm. We evaluate our proposed method using the International Symposium on Biomedical Imaging (ISBI) 2014 and 2015 challenge datasets. Results demonstrate that our method produces competitive accuracy compared to state-of-the-art methods.**

Keywords–Cervical cells; Shape prior; Low-rank optimization.

I. INTRODUCTION

Cervical cancer, which arises from abnormal growth of cervical cells, is the fourth most common cause of women's deaths from cancer. About 70% of these deaths occur in developing countries [1]. Fortunately, the incidence of cervical cancer has been significantly reduced by more than 50% over the past 30 years, mainly due to the widespread cervical cytology screening [2]. Until now, inspection of the cytology image has been done mainly manually, and the inspection process is labor intensive and suffers from inter and intra-observer variability. Hence, fully automatic computer diagnosis has attained great attention.

Cell segmentation is one of the critical steps in computer-aided diagnosis. In this step, individual cells are identified and their nuclei and cytoplasm are detected. The segmentation of overlapping cytoplasm is particularly challenging because of weak edges and inhomogeneous cytoplasm. Numerous attempts have been made to address the problem of overlapping cells. Lu et al. [3] proposed a joint optimization method using multiple level set functions, where different constraints are imposed including contour length, edge strength, cell shape and area of the overlapping regions. However, as pointed out by the authors, the traditional level set method used has issues with the initial segmentation and the re-initialization of the distance function. Hansang Lee et al. [4] proposed an improved algorithm by combining superpixels and cell-wise contour refinement through graph cuts, thus enhancing the oversegmentation method, which suffers from noisy and inaccurate boundaries, especially for overlapping regions. However, their contour refinement has the drawback of sometimes extracting boundaries along nucleus due to the nucleus having

the strongest boundary with large contrast. Phoulady et al. [5] proposed to use multiple images each with a different focal to delineate the cytoplasm corresponding to each nucleus. Since the edge features are not fully exploited for the cytoplasm segmentation, the performance of the algorithm is insufficient. Ushizima et al. [6] proposed a new algorithm for cytoplasm segmentation using a combination of nuclear narrow-band seeding, Voronoi diagram and graph-based region growing. They partition the cell mass into convex polygons such that only one nucleus is located in each Voronoi polygon. However, they assume that all segment lines have equal distances to the nearest nucleus - an assumption that is unrealistic for most overlapping cells. Nosrati et al. [7] proposed a variational approach for segmenting overlapping cells in Pap smear images. They exploit the regional prior, distance prior between a nucleus and its corresponding cytoplasm boundary, elliptical shape prior, and an overlap constraint that induces neighbouring cytoplasm to be excluded from one another. Compared to overlapping cytoplasm detection, nucleus detection is easier. Al-Kofahi et al. [8] proposed to segment cell nucleus using a combination of approaches. In their proposal, nuclear seed points are detected by multiscale Laplacian-of-Gaussian filtering constrained by adaptive scale selection and are used to perform an initial segmentation. A graph cuts algorithm incorporating alpha expansions and graph coloring is used to refine the segmentation. Plissiti et al. [9] used morphological analysis to detect the candidate nucleus centroids and refine the segmentation based on prior knowledge.

An outline of our method is summarized in Figure 1. Our approach consists of four main steps: nucleus localization, mass detection, coarse cytoplasm formation, and fine cytoplasm refinement. In step one, cell nuclei are detected by a Maximally Stable Extremal Regions (MSER) [10] detector and outliers which are tiny areas with low circularity are removed. In step two, the cell masses are extracted through triangle thresholding [11]. In step three, the coarse cytoplasm are generated via Simple Linear Iterative Clustering (SLIC) [12]. The superpixels are clustered based on distance measurement. In the final step, we apply an effective energy functional to refine the coarse cytoplasm. The energy functional consists of four terms: shape prior, regional prior, edge prior and regularized prior. The main contributions of our work are the shape prior construction, energy functional formulation, and the optimization strategy to minimize the energy functional.

In Section II, we give details on the construction of the shape prior using low rank optimization and the formulation of the energy terms that attract the active contour to different types of features. In Section III, we present a novel and effi-

cient algorithm that solves the energy minimization problem. In Section IV, we address the applicability of our method to cervical cell segmentation and present the performance of our method compared to three existing algorithms. Results demonstrate that our method generally shows competitive performance for the ISBI 2014 and 2015 public challenge datasets. Finally, Section V discusses some issues and provides a summary for our work.
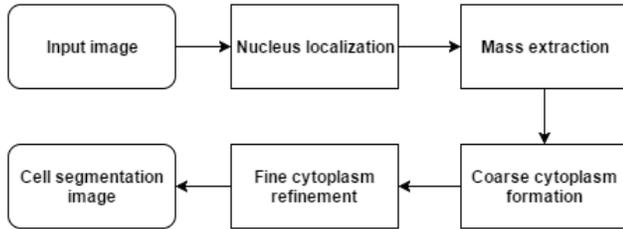
Figure 1. The workflow for overlapping cell segmentation.

II. MODEL SETUP AND PROBLEM FORMULATION

In this section, we show the construction of the shape prior using low rank optimization and the formulation of the energy terms that attract the active contour to different types of features.

A. Shape similarity

To cope with misleading features in images, shape prior is widely used to guide the segmentation process. It has been shown [13] [14] that for a matrix made up of parametric contours of shapes as column vectors, the rank of the matrix can act as a measure of the similarity of the shapes.

We use a vector $\mathbf{c} = [x_1, \ldots, x_n, y_1, \ldots, y_n]^T \in \mathbb{R}^{2n}$ as a parametric contour on the 2D plane, where (x_i, y_i) is a point on the contour. Assume there are k contours $\mathbf{c}_1$, ..., $\mathbf{c}_k$, and $\mathbf{c}_i (i \neq 1)$ is produced by $\mathbf{c}_1$ via affine transformation (i.e., translation, rotation and scaling). Then, the matrix made up of these contours has the following property

$$rank([\mathbf{c}_1, ..., \mathbf{c}_k]) \leq 6 \qquad (1)$$

This property is well-known but a quick proof is given here. $\mathbf{p}'$ and $\mathbf{p}$, the coordinates of a point before and after a shape change through affine transformation, follow the following relation: $\mathbf{p}' = \mathbf{A}\mathbf{p} + \mathbf{T}$. Since $\mathbf{c}_i$ is generated by $\mathbf{c}_1$ through affine transformation, we can express $\mathbf{c}_i$ as $\mathbf{c}_i = \mathbf{C}_1\mathbf{w}$ where

$$\mathbf{C}_1 = \begin{bmatrix} \mathbf{c}_x^1 & \mathbf{0} & \mathbf{c}_y^1 & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{c}_x^1 & \mathbf{0} & \mathbf{c}_y^1 & \mathbf{0} & \mathbf{1} \end{bmatrix},$$

$$\mathbf{w}_i = [a_{11}^i, a_{21}^i, a_{12}^i, a_{22}^i, t_1^i, t_2^i]^T.$$

Here, $\mathbf{c}_x^1 = [x_1, \ldots, x_n]^T$, $\mathbf{c}_y^1 = [y_1, \ldots, y_n]^T$, and a_{jk}^i and t_j^i denote the elements of $\mathbf{A}$ and $\mathbf{T}$. It is straightforward that $[\mathbf{c}_1, ..., \mathbf{c}_k] = \mathbf{C}_1[\mathbf{w}_1, ..., \mathbf{w}_k]$, and $\mathbf{w}_i$ has 6 rows. Therefore, $rank([\mathbf{c}_1, ..., \mathbf{c}_k]) \leq 6$.

Essentially, the rank of the matrix $[\mathbf{c}_1, ..., \mathbf{c}_k]$ represents the degrees of freedom of these shape changes. If we impose a low rank constraint on the matrix, shape changes such as, translation, rotation, and scaling will be allowed.

B. Active contour

In classical active contour models, there are generally two types of energy functional: edge-based [15] and region-based [16]. Next, we briefly review these models.

Let Ω, $\partial\Omega$ denote a bounded open set of $\mathbb{R}^2$ and the boundary of Ω respectively.

The edge-based snake model is to find the contour that produces:

$$\inf_{C(s)} E(C(s)) \qquad (2)$$

where

$$E(C(s)) = \mu \int_0^1 |C'(s)|^2 ds + \nu \int_0^1 |C''(s)| ds$$
$$- \lambda \int_0^1 |\nabla I_0(C(s))|^2 ds$$

is the energy functional.

Here, $C(s) \in \partial\Omega$ denotes a parameterized curve, I_0 denotes a given image, and μ, ν, λ are positive parameters. The first two terms in the energy functional are considered the internal energy, which controls the smoothness of the curve. The last term is considered the external energy, which attracts the curve to stop at the points of maxima $|\nabla I_0|$. Through minimizing $E(C(s))$, we are trying to locate the curve at the boundary of the object while keeping the curve itself as smooth as possible.

The region-based snake model is to find the contour that produces

$$\inf_{C(x,y), v_1, v_2} E(C(x, y), v_1, v_2) \qquad (3)$$

where

$$E(C(x, y), v_1, v_2) = \mu \cdot Length(C(x, y))$$
$$+ \nu \cdot Area(in(C(x, y)))$$
$$+ \lambda_1 \iint_{in(C(x,y))} |I_0(x, y) - v_1|^2 dxdy$$
$$+ \lambda_2 \iint_{out(C(x,y))} |I_0(x, y) - v_2|^2 dxdy$$

Here, $C(x, y) \in \Omega$ is a curve specified by a collection of points (x_i, y_i), I_0 denotes a given image, and v_1, v_2 respectively represent the average intensity of I_0 inside and outside $C(x, y)$. $\mu, \nu, \lambda_1, \lambda_2$ represent positive parameters. The first two terms are regularizing terms, which control the length of curve C and the area inside curve C. The last two terms are fitting terms. The basic idea of this model can be explained by a simple case. Assume the image I_0 is made up of two homogenous regions. The fitting terms are zero if and only if the curve is located at the boundary of the two regions.

C. Energy functional with shape prior

To fully exploit the edge-based and region-based active contour models, we propose to revise the energy functional and include a shape prior such that the problem becomes:

$$\inf_{C(x,y), v_1, v_2} E(C(x, y), v_1, v_2)$$
$$s.t. \quad rank(\Phi_t(C(x, y))) \leq N \qquad (4)$$

where,

$$E(C(x,y)) = \mu \iint_{on(C(x,y))} |\nabla C(x,y)|^2 dxdy$$
$$+ \nu \iint_{on(C(x,y))} |\triangle C(x,y)|^2 dxdy$$
$$+ \lambda_1 \iint_{on(C(x,y))} \frac{1}{1 + |\nabla G_\sigma * I_0|^2} dxdy$$
$$+ \lambda_2 \iint_{in(C(x,y))} |I_0 - v_1|^2 dxdy$$
$$+ \lambda_3 \iint_{out(C(x,y))} |I_0 - v_2|^2 dxdy$$

Here, $C(x,y)$ denotes a curve. For simplicity, we replace $C(x,y)$ by C for the rest of the paper. $\mu, \nu, \lambda_1, \lambda_2, \lambda_3$ are determined by validation on the ISBI datasets. v_1, v_2 respectively represent the average intensity of I_0 inside and outside C. To smooth the image, we use a 2-dimensional Gaussian kernel

$$G_\sigma = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{x^2 + y^2}{2\sigma^2}} .$$

Let $\Phi_t(C) = [C_{prior}, C_1^*, ...C_t^*, C]$. It contains the shape prior contour and a set of sub-optimal contour candidates and an unknown contour. C_{prior} can be arbitrary shape. Here, C_{prior} is a circular shape contour based on our task, and $C_1^*, ..., C_t^*$ are generated during the low rank minimization according to (5).

In general, the rank minimization problem is NP-hard. Therefore, relaxation is a common way to make the problem tractable. We use nuclear norm, which is proven to be the tightest convex surrogate of rank [17], to approximate the rank function, and relax the origin problem (4) to:

$$\inf_C E(C) + \lambda \|\Phi_t(C)\|_* \tag{5}$$

where $\|\Phi_t(C)\|_*$ denotes the nuclear norm of the matrix $\Phi_t(C)$. Suppose the rank of $\Phi_t(C)$ is r, then $\|\Phi_t\|_* = \sum_{i=1}^r \sigma_i$, where σ_i is the singular value of Φ_t. Hence, the nuclear norm can be viewed as the $l1$-norm of the vector of singular values.

III. THE PROPOSED MINIMIZATION ALGORITHM

Proximal gradient is one of the most popular methods for solving the problem in (5). To illustrate it, we consider a general unconstrained non-smooth non-convex optimization minimization problem:

$$\min_{\mathbf{X}} F(\mathbf{X}) + \lambda G(\mathbf{X}) \tag{6}$$

where F is a differentiable and G is a convex function. Further, F is assumed to be Lipschitz continuous: $\|\nabla F(\mathbf{X}_1) - \nabla F(\mathbf{X}_2)\| < \tau \|\mathbf{X}_1 - \mathbf{X}_2\|$. Therefore, F can be approximated locally as a quadratic function at $\mathbf{X}^k$ such that:

$$Q_\tau(\mathbf{X}, \mathbf{X}^k) = F(\mathbf{X}^k) + \langle \nabla F(\mathbf{X}^k), \mathbf{X} - \mathbf{X}^k \rangle$$
$$+ \frac{\tau}{2} \|\mathbf{X} - \mathbf{X}^k\|_F^2 + \lambda G(\mathbf{X})$$
$$= \frac{\tau}{2} \| \mathbf{X} - [\mathbf{X}^k - \frac{1}{\tau}\nabla F(\mathbf{X}^k)] \|_F^2 + F(\mathbf{X}^k) \tag{7}$$
$$+ \lambda G(\mathbf{X}) - \frac{1}{2\tau} \| \nabla F(\mathbf{X}^k) \|_F^2$$

where τ is a given positive parameter, $\langle \cdot, \cdot \rangle$ denotes the inner product, and $\| \cdot \|_F$ denotes the Frobenius norm. It is straightforward that (7) is a convex function of $\mathbf{X}$, and hence there exists a unique minimizer $\mathbf{X}^*$ such that:

$$\mathbf{X}^* = \arg \min_{\mathbf{X}} Q_\tau(\mathbf{X}, \mathbf{X^k})$$
$$= \arg \min_{\mathbf{X}} \frac{\tau}{2} \| \mathbf{X} - [\mathbf{X^k} - \frac{1}{\tau}\nabla F(\mathbf{X^k})] \|_F^2 \tag{8}$$
$$+ \lambda G(\mathbf{X}) + const$$

Now, we consider a special case of (8),

$$\min_{\mathbf{X}} \frac{\tau}{2} \|\mathbf{X} - \mathbf{S}\|_F^2 + \lambda \|\mathbf{X}\|_* \tag{9}$$

where $G(\mathbf{X}) = \|\mathbf{X}\|_*$. For convenience, we replace $\mathbf{X}^k - \frac{1}{\tau}\nabla F(\mathbf{X}^k)$ with $\mathbf{S} \subset \mathbb{R}^{m \times n}$ and ignore the constant term. Problem (9) has an analytical solution via Singular Value Decomposition (SVD) of $\mathbf{S}$, as has been proved in [18].

$$\mathbf{S} = \mathbf{U\Sigma V}^T, \mathbf{\Sigma} = Diag(\sigma) \tag{10}$$

Here, $\mathbf{U}, \mathbf{V}$ are respectively the left and right eigenvector matrices, and $\mathbf{\Sigma}$ is the diagonal matrix containing the singular values arranged in descending order. Let $\mathbf{x}_+ = max\{\mathbf{x}, \mathbf{0}\}$, and the solution $\mathbf{X}^*$ to (9) is given in closed-form as

$$\mathbf{X}^* = \mathbf{U}\text{Diag}((\sigma - \frac{\lambda}{\tau})_+)\mathbf{V}^T \tag{11}$$

In [19], usual Accelerated Proximal Gradient (APG) is extended to the general non-convex case by introducing a monitor variable that satisfies the sufficient descent property. The algorithm consists of the following steps:

$$\mathbf{Y}^k = \mathbf{X}^k + \frac{t^{k-1}}{t^k}(\mathbf{Z}^k - \mathbf{X}^k) + \frac{t^{k-1} - 1}{t^k}(\mathbf{X}^k - \mathbf{X}^{k-1}) \tag{12}$$

$$\mathbf{Z}^{k+1} = \text{prox}(\mathbf{Y}^k - \tau_y \nabla F(\mathbf{Y}^k)) \tag{13}$$

$$\mathbf{M}^{k+1} = \text{prox}(\mathbf{X}^k - \tau_x \nabla F(\mathbf{X}^k)) \tag{14}$$

$$t^{k+1} = \frac{1 + \sqrt{1 + 4(t^k)^2}}{2} \tag{15}$$

$$\mathbf{X}_{k+1} = \begin{cases} \mathbf{Z}_{k+1}, & \text{if} \quad F(\mathbf{Z}_{k+1}) \le F(\mathbf{M}_{k+1}) \\ \mathbf{M}_{k+1}, & \text{otherwise} \end{cases} \tag{16}$$

where τ_x, τ_y are constants satisfying $\tau_x \le \tau$ and $\tau_y \le \tau$. The proximal mapping is defined as $\text{prox}(\mathbf{X}) = \arg \min_{\mathbf{U}} G(\mathbf{X}) + \tau \|\mathbf{U} - \mathbf{X}\|^2$. The algorithm is an extension of monotone APG for the convex case. The difference lies in the introduction of the extra M matrix, which serves as the role of the monitor, and is used in the correction step during the X-update in (16).

To accelerate the convergence of the monotone APG algorithm, we propose to use a linesearch-like technique. In practice, it is desirable to set a smaller value for τ in (11) by performing a linesearch-like technique.

Now, we formally present the monotone accelerated proximal gradient with linesearch (APGL) algorithm, which is outlined as Algorithm 1.

Algorithm 1 Monotone APGL

Require: $\mathbf{X}^1 = \mathbf{X}^0, \mathbf{Z}^1 = \mathbf{Z}^0, t^1 = t^0, \tau^1 = \tau^0,$
$\eta \in (0, 1)$
for $k = 1$ to maximum iterations **do**
$\quad \mathbf{Y}^k = \mathbf{X}^k + \dfrac{t^{k-1}}{t^k}(\mathbf{Z}^k - \mathbf{X}^k) + \dfrac{t^{k-1}-1}{t^k}(\mathbf{X}^k - \mathbf{X}^{k-1})$
$\quad \widehat{\tau_1} = \eta\tau^{k-1}$
$\quad$ **for** $j = 1$ to maximum iterations **do**
$\quad\quad \mathbf{Z}^{k+1} = \mathrm{prox}(\mathbf{Y}^k - (\widehat{\tau_j})^{-1}\nabla F(\mathbf{Y}^k))$
$\quad\quad \mathbf{M}^{k+1} = \mathrm{prox}(\mathbf{X}^k - (\widehat{\tau_j})^{-1}\nabla F(\mathbf{X}^k))$
$\quad\quad \mathbf{X}^*_{\widehat{\tau_j}} = \begin{cases} \mathbf{Z}_{k+1}, & \text{if } F(\mathbf{Z}_{k+1}) \le F(\mathbf{M}_{k+1}) \\ \mathbf{M}_{k+1}, & \text{otherwise} \end{cases}$
$\quad\quad$ **if** $F(\mathbf{X}^*_{\widehat{\tau_j}}) \le Q_{\widehat{\tau_j}}(\mathbf{X}^*_{\widehat{\tau_j}}, \mathbf{Y}^k)$ **then**
$\quad\quad\quad \tau^k = \widehat{\tau_j},$ **break**
$\quad\quad$ **else**
$\quad\quad\quad \widehat{\tau_{j+1}} = min\left\{\eta^{-1}\widehat{\tau_j}, \tau^0\right\}$
$\quad\quad$ **end if**
$\quad$ **end for**
$\quad \mathbf{X}^{k+1} = \mathbf{X}^*_{\tau^k}$
$\quad t^{k+1} = \dfrac{1 + \sqrt{1 + 4(t^k)^2}}{2}$
$\quad$ **if** $\left\|\mathbf{X}^{k+1} - \mathbf{X}^k\right\|^2_F <$ tolerance **then**
$\quad\quad$ **return**
$\quad$ **end if**
end for

Figure 2. Monotone APGL.

Figure 3. The workflow and methods for overlapping cell segmentation.

IV. EXPERIMENT AND RESULTS

The outline of our cytoplasm segmentation method is summarized in Figure 2 and Figure 3. Now we describe our complete method in six steps: image smoothing, edge detection, mass extraction, nucleus localization, coarse cytoplasm estimation and fine cytoplasm refinement. An example input image is shown in Figure 3(a). First, we apply a bilateral filter to reduce noise and preserve edges. Result is shown in Figure 3(b). Then cell masses are extracted by triangle method [13], and result is shown in Figure 3(c). Edges are detected by SLIC [12], as shown in Figure 3(d). Cell nuclei are segmented by MSER, as shown in Figure 3(e). The coarse cytoplasm is generated by clustering the superpixels near the nucleus based on distance measurement. Finally, the fine cytoplasm is obtained by our proposed method, as shown in Figure 3(f).

We test our method on datasets released by recent ISBI challenges held in 2014 and 2015. The dataset consists of 16 EDF real cervical cytology images and 945 sythetic images. The 810 synthetic images are used for testing in the quantitative evaluation. The 2015 dataset consists of a collection of 17 multi-layer cervical cell volumes, from which 8 will be used for training and 9 for testing. They are realistic images with overlapping cells and poor contrast. We evaluate our method quantitatively by a set of four metrics. Specifically, first, the Dice coefficient(DC) is measured as $DC = \dfrac{2|A \cap B|}{|A| + |B|}$. For individual cytoplasm segmentation, the Dice coefficient is computed over valid cytoplasm segmentation. A valid segmentation is defined as $DC > 0.7$. Second, the false negative rate (FN_o) is calculated as the proportion of segmentation with $DC \le 0.7$. In addition, for valid segmentation, the true positive rate (TP_p) and false positive rate (FP_p) are also
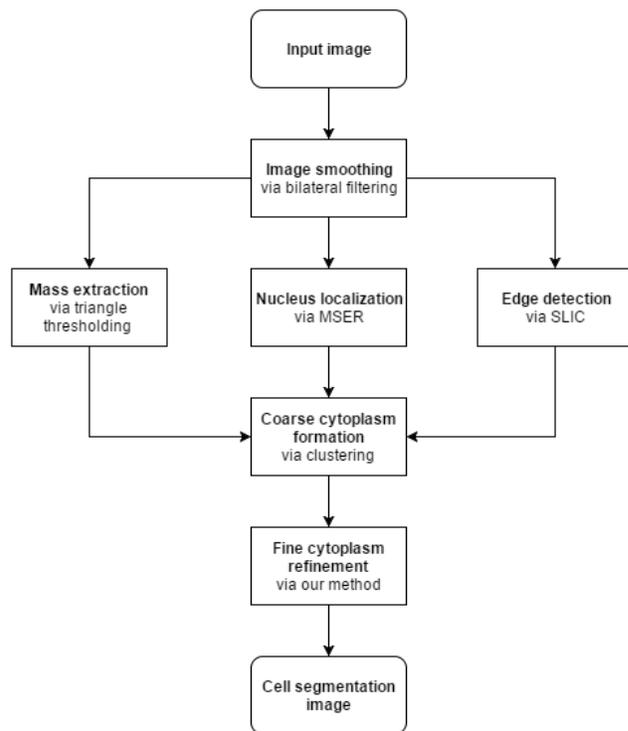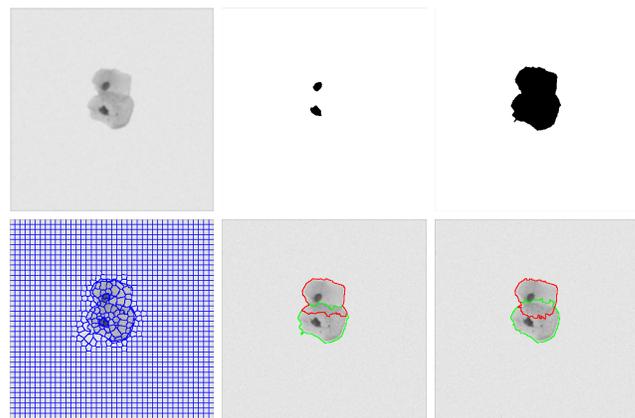
Figure 4. The images of workflow for cytoplasm segmentation. (a) Filtered image (b) Nucleus image (c) Mass image (d) Edge image (e) Coarse cytoplasm image (f) Fine cytoplasm image.

presented as the third and fourth metrics.

We compare our method with existing methods [21]. Specifically, [6], [7] are the winners of ISBI 2014 challenge and [5], [20] are the the winners of ISBI 2015 challenge. In Table I, our proposed method achieves the lowest object-based false negative rate on both datasets. Especially, the object-based false negative rate has been significantly reduced. This means our method generates many more instances of valid segmentation than other existing methods.

Figure 4 displays the segmentation results under different degrees of overlapping cells. As shown in the first row, we divide the overlapping cells into three categories: non-overlapping cells, pair overlapping cells, and multiple overlapping cells. The second row shows the results after the coarse contour formation, and the third row displays the results after
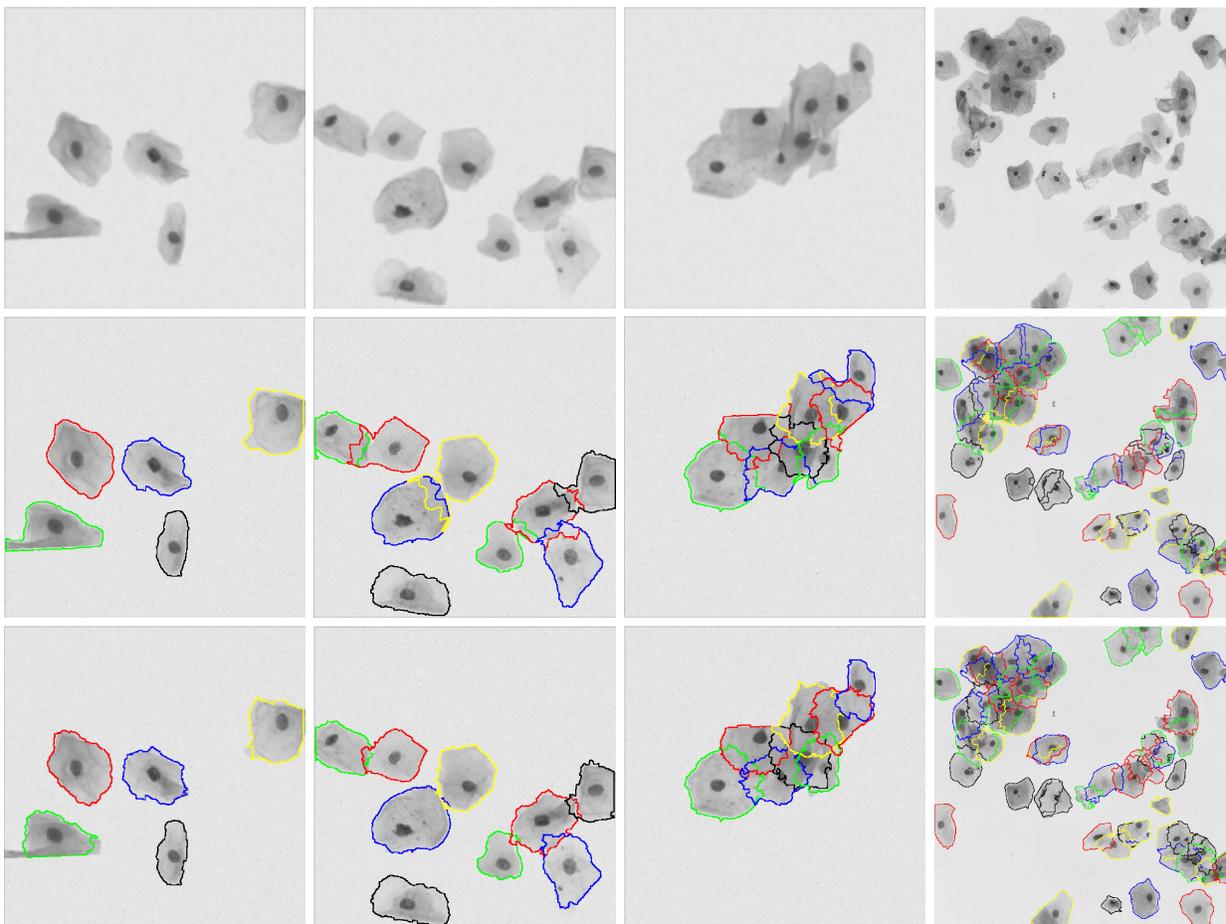
Figure 5. Results of our cytoplasm segmentation results. The top row shows original images. The middle row shows the output after coarse cytoplasm formation. The bottom row shows results of fine cytoplasm refinement.

TABLE I.   EVALUATION AND COMPARISON OF OUR CYTOPLASM SEGMENTATION RESULTS WITH THREE STATE-OF-THE-ART METHODS.

Datasets	Method	FNo	TPp	FPp	DC
ISBI 2014	Ushizima [6]	0.267	0.841	**0.002**	0.872
	Nosrati [7]	0.111	0.875	0.004	0.871
	Lu [3]	0.316	**0.905**	0.004	0.893
	Lee [4]	0.137	0.882	0.002	**0.897**
	Ours	**0.098**	0.901	0.004	0.875
ISBI 2015	Phoulady [5]	0.408	**0.927**	0.003	0.831
	Ramalho [20]	0.501	0.899	0.002	0.856
	Lee [4]	0.434	0.877	**0.001**	**0.879**
	Ours	**0.285**	0.861	**0.001**	0.856

fine cytoplasm refinement. Our method successfully segments all cell cytoplasm in the first two columns with a Dice coefficient around 0.95 for both cases. Our method misses one cytoplasm in the last column due to the severe overlap and weak contrast. The Dice coefficient of the last column is above 0.85, which is still an acceptable result. The segmentation results show a good delineation of the overlapping cytoplasm, and indicate the strong ability of our proposed method in extracting the cell cytoplasm.

In addition, the average running time of our method is about 2.5 seconds for each cell. Our code is written in Matlab and run in Ubuntu. The configuration is a notebook with Intel core i7 and 16 GB RAM.

V.   DISCUSSION AND CONCLUSION

In the experimental results, our proposed method in general achieves better performance than the three existing methods. Especially, we observe that the object-based false negative rate has been greatly reduced. We attribute the improvement to the cytoplasm refinement via our proposed method. Since we achieve the highest true positive rate, it is reasonable that our false positive rate is slightly higher than other methods. According to the qualitative results in Figure 4, our approach successfully segments the individual cytoplasm for most cells under different degrees of overlap. Examining the results in Figure 4, we can see that the shape prior plays an important role in fine cytoplasm refinement. The shape prior makes the cytoplasm more elliptical. It is the key factor for improving the false negative rate. But, a point that needs to be addressed

is that, as shown in the first column of Figure 4, the fine cytoplasm refinement actually degrades the result for the cell in green because the shape prior rejects the non-elliptical protrusion on the left-side of the cell.

In conclusion, we have proposed a novel fully automatic segmentation method for overlapping cells. The main contributions of our work are the shape prior construction, energy functional formulation, and the optimization strategy to minimize the energy functional. The shape prior is modeled using low rank optimization to deal with the weak edges and low contrast in overlapping cells. The terms of the energy functional are designed to exploit both the edge and regional features. The monotone APGL algorithm presented solves our non-smooth non-convex problem efficiently. We test our proposed method on the ISBI 2014 and 2015 challenge datasets, and our method shows competitive accuracy compared with existing state-of-the-art methods.

REFERENCES

[1] B. Stewart, C. P. Wild *et al.*, "World cancer report 2014," *World*, 2016.

[2] N. Behtash and N. Mehrdad, "Cervical cancer: screening and prevention," *Asian Pac J Cancer Prev*, vol. 7, no. 4, pp. 683–6, 2006.

[3] Z. Lu, G. Carneiro, and A. P. Bradley, "An improved joint optimization of multiple level set functions for the segmentation of overlapping cervical cells," *IEEE Transactions on Image Processing*, vol. 24, no. 4, pp. 1261–1272, 2015.

[4] H. Lee and J. Kim, "Segmentation of overlapping cervical cells in microscopic images with superpixel partitioning and cell-wise contour refinement," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 63–69.

[5] H. A. Phoulady, D. B. Goldgof, L. O. Hall, and P. R. Mouton, "An approach for overlapping cell segmentation in multi-layer cervical cell volumes," *The Second Overlapping Cervical Cytology Image Segmentation Challenge-IEEE ISBI*, 2015.

[6] D. M. Ushizima, A. G. Bianchi, and C. M. Carneiro, "Segmentation of subcellular compartments combining superpixel representation with voronoi diagrams," *Overlapping Cervical Cytology Image Segmentation Challenge-IEEE ISBI*, pp. 1–2, 2014.

[7] M. S. Nosrati and G. Hamarneh, "A variational approach for overlapping cell segmentation," *ISBI Overlapping Cervical Cytology Image Segmentation Challenge*, pp. 1–2, 2014.

[8] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam, "Improved automatic detection and segmentation of cell nuclei in histopathology images," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 841–852, 2010.

[9] M. Plissiti, A. Charchanti, O. Krikoni, and D. Fotiadis, "Automated segmentation of cell nuclei in pap smear images," in *Proc. IEEE International Special Topic Conference on Information Technology in Biomedicine*. Citeseer, 2006, pp. 26–28.

[10] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and vision computing*, vol. 22, no. 10, pp. 761–767, 2004.

[11] G. Zack, W. Rogers, and S. Latt, "Automatic measurement of sister chromatid exchange frequency." *Journal of Histochemistry & Cytochemistry*, vol. 25, no. 7, pp. 741–753, 1977.

[12] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[13] X. Zhou, X. Huang, J. S. Duncan, and W. Yu, "Active contours with group similarity," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2969–2976.

[14] X. Zhou, C. Yang, H. Zhao, and W. Yu, "Low-rank modeling and its applications in image analysis," *ACM Computing Surveys (CSUR)*, vol. 47, no. 2, p. 36, 2015.

[15] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1988.

[16] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Transactions on image processing*, vol. 10, no. 2, pp. 266–277, 2001.

[17] M. Fazel, "Matrix rank minimization with applications," Ph.D. dissertation, PhD thesis, Stanford University, 2002.

[18] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[19] H. Li and Z. Lin, "Accelerated proximal gradient methods for non-convex programming," in *Advances in Neural Information Processing Systems*, 2015, pp. 379–387.

[20] G. L. Ramalho, D. S. Ferreira, A. G. Bianchi, C. M. Carneiro, F. N. Medeiros, and D. M. Ushizima, "Cell reconstruction under voronoi and enclosing ellipses from 3d microscopy," in *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2015.

[21] Z. Lu, G. Carneiro, A. Bradley, D. Ushizima, M. S. Nosrati, A. Bianchi, C. Carneiro, and G. Hamarneh, "Evaluation of three algorithms for the segmentation of overlapping cervical cells," *IEEE journal of biomedical and health informatics*, 2016.