SIGBio Record

Newsletter of the SIGBio
ACM Special Interest Group

Volume 7, Issue 3, Dec 2017 ISSN 2159-1210

Notice to Contributing Authors to SIG Newsletters

By submitting your article for distribution in this Special Interest Group publication, you hereby grant to ACM the following non-exclusive, perpetual, worldwide rights:

• to publish in print on condition of acceptance by the editor

• to digitize and post your article in the electronic version of this publication

• to include the article in the ACM Digital Library and in any Digital Library related services

• to allow users to make a personal copy of the article for noncommercial, educational or research purposes

However, as a contributing author, you retain copyright to your article and ACM will refer requests for republication directly to you.

SIGBIO Record - Submission Guidelines

Submission categories

Submissions to the newsletter can be either on a special issue topic or on topics of general interest to the SIGBIO community.

These can be in any one of the following categories:

- Survey/tutorial articles (short) on important topics.
- Topical articles on problems and challenges
- Well-articulated position papers.
- Review articles of technical books, products and .
- Reviews/summaries from conferences, panels and special meetings within 1 to 4 pages [1500-2500 words]
- Book reviews and reports on relevant published technical books
- PhD dissertation abstracts not exceeding 10 pages
- Calls and announcements for conferences and journals not exceeding 1 page
- News items on the order of 1-3 paragraphs

Brief announcements Announcements not exceeding 5 lines in length can simply be sent as ASCII text to the
editors by e-mail. SIGBIO Record publishes announcements that are submitted as is without review.

Announcements cannot be advertisements and should be of general interest to the wider community. The Editor reserves the right to reject any requests for announcements at his discretion.

Authors are invited to submit original research papers or review papers in all areas of bioinformatics and computational biology. The papers published in SIGBioinformatics Record will be archived in ACM Digital Library. Papers should follow the ACM format, and there is no page limitation.

http://www.acm.org/sigs/publications/proceedings-templates

Submissions should be made via email to one of the two editors Pierangelo Veltri (veltri@unicz.it) *(University Magna Graecia of Catanzaro, Italy)- Pietro Hiram Guzzi (hguzzi@unicz.it),or to ACM SIGBio email acmsigbiorecord@gmail.com*

2017 ACM Fellows

Dear SIGBio members,

I am delighted to share the good news that two prominent members of our community are honored as 2017 ACM Fellows (https://www.acm.org/media-center/2017/december/fellows-2017).

Prof. Aidong Zhang, SUNY Distinguished Professor at SUNY Buffalo and Founding Chair of our society ACM SIGBio, who is currently serving as a program director at NSF, is elevated to ACM Fellow for her *contributions to bioinformatics and data mining*. She is also the current editor-in-chief of the IEEE/ACM Transactions on Computational Biology and Bioinformatics, steering committee co-chair of the ACM BCB Conference, and continues to serve SIGBio as chair of its advisory board. She authored books on gene expression analysis (https://www.amazon.com/Advanced-Analysis-Gene-Expression-Microarray/dp/B0075LS6CQ/) and protein interaction networks (https://www.amazon.com/Protein-Interaction-Networks-Computational-2009-04-06/dp/B01FIZ2LCG/).

Prof. Dan Gusfield, Distinguished Professor at UC Davis and author of an early and highly popular textbook on computational biology (https://www.amazon.com/Algorithms-Strings-Trees-Sequences-Computational/dp/0521585198/), is elevated to ACM Fellow for his *contributions to combinatorial optimization and algorithmic computational biology*. He also served as the founding editor-in-chief of the IEEE/ACM Transactions on Computational Biology and Bioinformatics. He can also tell us a thing or two about stable marriages :(https://www.amazon.com/Stable-Marriage-Problem-Algorithms-Foundations/dp/0262515520/).

Sincerely,
Srinivas Aluru
SIGBio Chair

Topic: Topical articles on problems and challenges

THE UNCAP EXPERIENCE IN DEVELOPING ICT FOR HEALTH: "HTA BY DESIGN"

Anzivino S*, Tessarolo F*, **, Morganti E**, Conti G[o], Nollo G*, **

* Healthcare Research and Innovation Program - Health Technology Assessment, PAT-FBK, Trento, Italy

** Department of Industrial Engineering & BIOTech, University of Trento, Italy

[o]Trilogis Srl, Rovereto (TN), Italy

ABSTRACT

Dementia is one of the most relevant chronic diseases affecting the ageing population. The elder with mild or moderate cognitive impairment (MMCI) suffers of progressive cognitive decline with increasing difficulties in performing activities of daily living. Information and Communication Technology (ICT) for Healthcare can provide solutions to relief the caregivers' burden and to support the elder in maintaining dignity and independence. The H2020 project aimed at developing and testing a bundle of hardware and software technologies able to fit the individual needs of the elder with MMCI and his/her formal and informal caregivers. The revolutionary design of UNCAP technologies and services required the development of a new paradigm for assessing their impact on the care system. Health Technology Assessment was the suitable instrument for multidimensional evaluation (safety, effectiveness, costs, impact). These tools was further refined and applied to UNCAP technologies by integrating HTA in all planning and development phase of the product, thus realizing the "HTA by design". According to this new paradigm, all dimensions of analysis were taken into account starting from the conceptualization of UNCAP solution and requirement elicitation, to the prototype development and testing at pilot sites.

Dementia in the EU aging population

The world's population is ageing rapidly with an estimation of 1 in 5 people over 65 years old by 2030 compared to 1 in 10 today. Due to chronic age-related illnesses, many progressively lose their autonomy and become more dependent on others, finally reaching the stage when they need round-the-clock care from their family members or caregivers. One of the most important chronic diseases that affect the ageing population is dementia. It accounts for 4.1% of total disease burden among people aged over 60 years and 40% of people older than 85. The number of people affected by this disease is increasing exponentially with an estimation of 35.6 million people with dementia in 2010, and numbers nearly doubling every 20 years [Prince, 2009].

According to the Global Deterioration Scale (GDS) [Reisberg, 1988], cognitive and functional abilities are categorized into 7 stages, ranging from no cognitive decline in the first stage to very severe cognitive decline in the seventh stage. Stage 5 denotes the point where it becomes difficult for the patient to live independently and assistance is needed from his/her family and/or caregivers. During stages 3 to 5 (mild and moderate cognitive impairments, MMCI), the elder suffers progressive cognitive decline and experience increasing difficulties in performing activities of daily living (ADLs). In some instances, MMCI elders may understand what they are supposed to be doing but they may not understand the instructions, or forget them midway through a task. They may also fail to recognize objects for what they are (agnosia) or to know how to execute learned tasks (apraxia). This means that the caregivers have to be present to help patients during their activities, and over time, increase the support they provide as the disease evolves [Aloulou, 2013].

Indeed, older adults, including people with MMCI, desire to remain in their homes as they age [CDC, 2013], creating significant challenges to manage the needs of increasing care while living at home. Since the 1980s, technology has been investigated as a possible support for the so called "aging in place" [Cook, 2009]. The technology advancesbring new opportunities to reduce both the burden of caregiving and the need for premature nursing home placement, due to family caregivers no longer being able to meet care demands.

People with MMCI have a risk of hospitalizations and nursing home admissions triplicated compared to older adults with other conditions [Bossen 2015].

Family caregivers experience high levels of stress, burden, and role captivity that lead to negative physical, psychological, social, and spiritual outcomes [Zarit 2006; Brodaty, 2002; Schulz, 2008]. Caregivers of people with MMCI must cope with their loved one's progressive memory loss, self-care impairment, and communication breakdown. Caregiving stress, strain, and burden contribute to negative physical and mental health outcomes that include depression, insomnia, and psychotropic medication use, with notable increases in caregiver morbidity and mortality [Monin 2009]. Caregivers separated by distance face unique challenges as they manage caregiving from afar. They may worry about their family member's safety and security, medication schedules, wandering, and need for information and socialization. The distant

caregiver may be totally unaware of the needs of their family member, placing further burden on the onsite caregiver(s) [Bossen 2015].

UNCAP: personalized technology on individual needs

Enhancing the well-being of people with MMCI and of their caregivers is a complex and evolving task. Information and Communication Technology (ICT) for Healthcare (e-Health) can effectively play an important role, provide solutions to relief the burden on caregivers, and support the individuals, with MMCI or other impairments, in maintaining dignity and independence while they age.

UNCAP ("Ubiquitous iNteroperable Care for Ageing People", GA 643555) is an European project that fosters a modern non-pharmacological approach as an appropriate initial strategy in the support and care of individuals with MMCI with the aim of improving users' quality of life.

The project involved 23 partners (including several pilot user partners) from 9 European countries (IT, UK, SI, RO, EL, DE, SE, ES, MK) with multidisciplinary backgrounds and competences.

UNCAP has the aim to create and test a bundle of hardware and software technologies, customizable and flexible enough to adapt to the needs of the elder with MMCI and his/her formal and informal caregivers. In particular UNCAP technology provides:

- Monitoring systems for:
  - Monitoring of physiological parameters
  - Monitoring of physical activity levels
  - Monitoring of activities of daily living
  - Falls detection
- Ambient Assisted Living for:
  - Communications
  - Emergency call
  - Geofencing
- Tracking and wayfinding for:
  - Falls prevention
- Cognitive stimulation with:
  - Serious games
- Cognitive aids as:
  - Reminder
  - Instructions.

The complexity of toolsconsisting of both hardware and software components and incorporating customizable tools and services for assistive living (fall prevention, cognitive aids), monitoring (tracking and geo fencing), diagnostics (remote sensing, physiological parameters monitoring)

and therapy (serious games, cognitive stimulation, exergames),as for UNCAP bundle, requires the development of new paradigms of performance and safety assessment,as well as its impact on the healthcare system.

Since UNCAP technologies were conceived for treating, caringand alleviating a disease or impairment, UNCAP has been considered, from the beginning, as a Medical Device. According to this intents of use, the design, development and assessment were carry out in compliance with the European Directives and Regulations on Medical Devices (93/42/CEE andsubsequent amendments).

In order to comply the forthcoming Regulation on Medical Device (2017/745 EU), data on effect should be provided in addition to safety for obtaining CE mark. Being UNCAP composed of innovative technologies, device safety and effect could not be extrapolated from literature and new clinical investigations should be realized. Moreover, a more comprehensive evaluation, covering also the economic and social impact for the introduction of a radically new technology is advised.

Health Technology Assessment by design

To fit this purpose, Health Technology Assessment (HTA) is a scientific methodology able to evaluate in a more comprehensive way several different dimensions including safety, effectiveness, costs, impacts and more. An European Project (EUnetHTA) delivered a reference framework for the HTA methodologies, called "HTA Core Model" [http://www.eunethta.eu/hta-core- model], aiming at the universalization of the elements of an HTA evaluation. However, ICT applications for health present specific characteristics, in terms of reliability, accuracy, etc. compared with other medical devices, making the traditional HTA approach, and also the "HTA Core Model", unsuitable and not easily applicable.

More recently, a new goal was reached in Telemedicine, a branch of e-Health, defining the "Model for Assessment of Telemedicine" (MAST) [http://www.mast-model.info/] delivered by the MethoTelemed European Project. MAST re-adjusted the "HTA Core Model", identifying the following seven dimensions for the analysis of Telemedicine technologies:

- Health's problem and use of technology;
- Safety;
- Clinical effectiveness;
- Patient's perspective;
- Economical aspects;
- Organizational aspects;
- Socio-cultural, ethical and legal aspects.

MAST assessment methods wereapplied for UNCAP technologies, integrating the methodology into the planning and development phase of the product. So that, all MAST dimensions were considered since the conception of the device (not limiting just to safety and performance evaluation) thus moving towards the concept of “HTA by design”. Indeed, the MAST approach was used to write the proposal, to extract the system requirements, to realize the prototypes and to test UNCAP technology in pilot studies.

The UNCAP multicentric trial

Following this multi-dimensional evaluation approach, a multicenter clinical investigation was designed for assessing improvements in the quality of life of all users (primary and secondary users) and the impact on the use of resources for care. The objective of the investigation was also to assess safety and usability of UNCAP in responding to the needs of elderly people with mild and moderate cognitive impairment as well as evaluating primary and secondary users’ acceptance and satisfaction.

Quantitative and semi-quantitative evaluation tools and methods, including the collection of quantitative, qualitative and narrative informationwere retrieved from literature (validated questionnaires) or specifically developed (structured questionnaires). Devoted questionnaires were conceived an applied to elicit users’ needs, identifying the technology use-cases fitting best to theneed and finally configuring the optimal UNCAP technology user-by user.UNCAP modularity allowed the customization of the bundle features according to the actual care setting and the users’ needs (both primary and secondary users’ needs).

Six Pilot sites in Italy were involved in this clinical investigation. Five more pilot sites were involved in testing UNCAP technology across Europe. Each of the pilots implemented a specific set of UNCAP features chosen according to the specific application scenario, environment and users’ needs. According to this, UNCAP was tested for caring elderly people with MMCI in both long term care facilities as an additional care device (residential care setting) and at the primary user home for providing home care services (home care setting).

The assessment plan dealt with the complexity of defining clear and comprehensive end-points and outcome variables, preserving the possibility of configuring the UNCAP technology according to the specific needs elicited by the user.Indeed,the modularity and adaptability of UNCAP in different scenarios reflects the complexity of the clinical investigation to assess UNCAP usability and safety, users’ acceptance, satisfaction and quality of life. However, all pilot sites shared the same research questions, with a common set of primary and secondary outcomes and evaluation tools.

A complex set of assessment tools, some of them borrowed from HTA methodology, allowed to explore a wide range of dimensions and to extract common indicators and outcomes (Table 1),making them available since the beginning of the UNCAP development in order to realize the “HTA by design” concept.

The investigation wasimplemented as a multicenter randomized controlled prospective parallel multicenter study.

At all times, data sharing and analysis has respected the dignity, privacy and confidentiality of individual's personal information.

The study was approved by all local ethic committees and it is currently ongoing, approaching the final phases of execution in all pilot sites.

Project results will be made public at the end of the project on the project web-site (www.uncap.eu).

Table 1: Dimensions tools and indicators used for assessing UNCAP technology according to the concept of "HTA by design"

Dimension	**Assessment tool**	**Indicators/outcomes**
Safety	• Systematic review • Risk analysis • Adverse event reporting form (according to MEDDEV 2.7/3)	• Number of adverse events • Number of device-related adverse events • Number of Serious Adverse Events
Clinical effectiveness	• Systematic review • INTERRAI assessment tools(http://www.interrai.org/) • Validated questionnaires (QoL, FES-I)	• Personal Health Profile (PHP) key from ATL@NTE (http://www.sistematlante.it/) • Perceived risk of falls • Quality of Life of the primary user • Quality of life of the informal caregiver
Patient's perspective	• Structured questionnaires • Users' Interviews	• Perceived usability • User acceptance • User satisfaction
Economical aspects	• Costs analysis • Business plan	• Return of Investment • Expenditures - materials, personnel, services, communication, etc. • Received public investment per year • Amounts of public investments per year • Amounts of private investments per year • Amounts of changed paid-in-capital or equity
Organizational aspects	• Questionnaire	• Number of medical examinations by general practitioner • Number of medical examinations by other physicians • Number of referrals to the emergency department • Number of hours per month spent by nurses caring for participant • Number of hours per month spent by formal and informal caregivers in taking care or patient • Number of days off work for family members • Number of technical interventions for device malfunction
Social aspects	• Structured questionnaires	• Formal and informal caregivers satisfaction • Informal caregiver quality of life

	• Validated questionnaires (SQLC)	
Legal aspects	• International regulation analysis on privacy and security • International regulation analysis on medical device	• Data management plan • Privacy by design • Compliance with medical devices regulations by design
Ethical aspects	• International regulation analysis on ethics • Ethical committee interrogation	• Approval by competent ethical committees

Bibliography

- Aloulou H, Mokhtari M, Tiberghien T, et al. Deployment of assistive living technology in a nursing home environment: methods and lessons learned. BMC Medical Informatics and Decision Making. 2013;13:42. doi:10.1186/1472-6947-13-42.
- Bossen AL, Kim H, Williams KN, Steinhoff AE, Strieker M. Emerging roles for telemedicine and smart technologies in dementia care. Smart homecare technology and telehealth. 2015;3:49-57. doi:10.2147/SHTT.S59500.
- Brodaty H, Green A. Who cares for the carer? The often forgotten patient. AustFam Physician. 2002; 31(9):833.
- Centers for Disesase Control and Prevention. [Accessed January 7, 2015] The State of Aging and Health in America. 2013. Available from: http://www.cdc.gov/aging/help/dph-aging/state-aging-health.html
- Cook DJ, Augusto JC, Jakkula VR. Ambient intelligence: technologies, applications, and opportunities. Pervasive Mob Comput. 2009; 5(4):277–298.
- Monin J, Schulz R. Interpersonal effects of suffering in older adult caregiver relationships. Psychol Aging. 2009; 24(3):681–695.
- Prince M, Jackson J, International AD: World Alzheimer Report 2009. In Alzheimer's Disease International; 2009. [http://www.alz.co.uk/research/files/WorldAlzheimerReport.pdf]
- Reisberg B, Ferris S, De Leon M, Crook T, et al: Global Deterioration Scale (GDS). *Psychopharmacol Bull* 1988, 24(4):661
- Schulz RS, Sherwood PR. Physical and mental health effects of family caregiving. Am J Nurs. 2008; 108(Suppl 9):23–27.
- Zarit, S. Assessment of Family Caregivers: A research perspective. San Francisco; CA: Family Caregiver Alliance; 2006.

Harnessing Mass Spectra Data Using KNN Principle; Diagnosing Alzheimer's Disease

Destiny E. O. Anyaiwe[1*], George D. Wilson[2], Timothy J. Geddes[3] Gautam B. Singh[4]

[1,4]Department of Computer Science and Engineering, Oakland University, MI, USA
[2,3]William Beaumont Hospital, Royal Oak, MI, USA
[*]**Corresponding Author:** Destiny Anyaiwe, Department of Computer Science and Engineering, Oakland University, Rochester,MI, 48309, USA; Tel:+1248-370-2129, oanyaiwe@oakland.edu

Abstract

A high level of expertise, rigorous algorithms and methods are needed to adequately mine and harness Mass Spectrometer generated data due to its unique nature and structure. Hitherto, peptide ions are matched with theoretical results and/or public databases in order to identify expressed proteins in analyzed protein source samples, but this is done on a spectrum by spectrum basis. In this study, we present a mechanism that extends the principle of K-nearest neighbor algorithm for mining pools of mass spectrometer saliva data towards discovering and characterizing patterns for diagnosing Alzheimer's disease. The methodology discusses feature selection by correlation matrix, matrix to vector decomposition, an extension of *Euclidean* distance formula, and successfully classifies donor samples into the three stages of Alzheimer's disease with over 85% accuracy without collaborating clinical records.

Keywords: Alzheimer's Disease Diagnosis; Feature matrix; Jackknifing; Mass Spectra Data

1 Introduction

In the US, from 2000 to 2013, while deaths from other diseases declined significantly [1], but that of *Alzheimer's* disease (AD) increased by 71%. The disease currently affects over five million people in the US, and the number is expected to grow to 16 million by 2050; afflicting one in nine people over the age of 65, and one in three people over the age of 85.

The unfortunate realities today is that 1) there are no cures for the disease and 2) early diagnosis is key. This is further complicated with the lack of clinical diagnostic tools. The clinical practice for diagnosing AD today is patients follow up system, where patient's cognitive abilities are judged based on the state of their memory through protracted 'Q & A' sessions. The practice is counter productive and non scientific, the follow up can be for many months (during which, for instance, mild cognitive impairment cases could degenerate to full blown dementia). Furthermore, the practice most often than not leads to inconclusive diagnosis and results or clinical notes achieved through them are not generalizable.

The building of a classification model using a pool of Mass Spectrometer Surface Enhanced Laser Desorption/Ionization (SELDI) time-of-flight *saliva* data is the aim of this study, it extends principle of K-Nearest Neighbor (KNN) Algorithm. In addition, it develop additional use for such data and closes the gap between protein expressions as detailed in Mass Spectrometer generated data and the diagnoses of Alzheimer's disease.

The next section gives the literature review. Section 3 describes our methodology. Discussion of our results and observations are given in Section 4 and possible areas of future works and conclusion are highlighted in Section 5.

2 Literature

The principle of KNN is a non parametric algorithm that adopts appropriate distance functions to induce measure on the location of instances of the train data-set from a test data-point. It then classify the test data based on the label of the data points most closest it. KNN is easy to implement, modify, extend and adept for data-sets with 3-to-4 classes, [2].

Conceptualizing individual objects (e.g. genotypes, vectors) as elements existing in a multidimensional spaces aids the application of geometric classification techniques. It makes the creation of homogeneous groups by building data from the structure of correlated groups in the multidimensional space possible [3]. Leping et al. in [4] explored KNN with Genetic Algorithms as an approach for the generation of predictive gene subsets.

In [5], multi-labelling based on identifying the KNNs of training set in instances of test set was presented, it further showed how such exercise can be used to predict yeast gene functionality, assign labels to unseen images in natural scene classification problems, as well as solving web page automated categorization problems. Similar idea was presented in [6] for image recognition.

Structural proteomics was studied in [7]. The study achieved grouping and predicting of new proteins based on structure alignments of the distance matrices obtained by 2D representation of protein's tertiary structures.

Sundry studies about phylogenetic tree construction, nodes connection in social and biological network systems utilize different forms of distance functions, ([8], [9]). The results of such studies can be extended for classification or predicting purposes if supplemented with the generalization principles of KNN.

This study utilized, Jackknife sampling procedure to constitute the training and corresponding test data-sets. The importance and reasons as to why and when Jackknife technique can be used, were presented by [10], and the method was applied in feature selection and classification by [11].

For a protein source sample (PSS) e.g. saliva, SELDI generates a spectrum with hundreds of peaks that are typically categorized by the following properties; the mass-to-charge ratio (m/z), time-of-flight (TOF), TOFIntensity, Substance mass, ionCharge, ionMass, signal-to-noise ratio and peak type, which are used for further investigation of ions with respect to investigator's objectives. The investigation of peptide peaks usually begins with detecting the set of peaks that are *'differentially expressed'* in the spectrum after baseline subtraction has been done using statistical methods or thresholding.

To identify the protein constituent of a PSS, the molecular weights, other features of ions and their chemical properties (e.g. which chemical surface it binds to preferentially on the ProteinChip) are matched with public databases or theoretical results. Definitive identification of the peak is then carried out using other algorithms (e.g. proteinProphet; [12]). Other uses of SELDI data are in the area of determining molecular formulas, protein curating and identification, and protein bio-marker discovery [13], personalized medicine, drug design and drug production ([14],[15]).

3 Methodology

The data-set used for this study was obtained from the BioBank of Beaumont Reference Laboratory, Michigan, USA. It was the output of a SELDI-TOF discovery proteomics laboratory experiment carried out on saliva to assess differential protein expressions for the purpose of identifying protein biomarkers for Alzheimer's disease (AD). Three populations of patients whose stages were known a priori were studied; age-matched controls without any evidence of dementia (CON), patients with mild cognitive impairment (MCI) and patients with clinical symptoms of Alzheimer's disease (tAD).

The classification model presented in this paper characterizes assay results. Assay results are matrices made up of by tens to hundreds of differentially expressed ions. This matrix structure makes feature selection, pattern recognition and building of classification models impossible, since direct application of traditional Supervised/Unsupervised machine learning algorithms fails. This is so because, those algorithms accepts feature vectors as inputs and discriminate data points on a *point-to-point* basis whereas the data type under consideration are matrices.

3.1 Data Set & Feature Selection

The 'uniqueness' of the data-set is due to the structure of SELDI analysis results. Each result or data point is a matrix, having a collection of tens/hundreds of observed ions. Matrix (R) below is a typical representation of such results.

$$R_k^S = \begin{bmatrix} m_1 & T_1 & I_1 & S_1 & C_1 & M_1 & N_1 & P \\ m_2 & T_2 & I_2 & S_2 & C_2 & M_2 & N_2 & P \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ m_n & T_n & I_n & S_n & C_n & M_n & N_n & P \end{bmatrix} \tag{1}$$

where[1] $k = \{1, 2, \cdots, 20\}$ indexes the total number of results in each disease stage (S). Elements of R are arranged in an ascending order of magnitude of m/z values.

Mass spectrometer calibrates feature values using different scales. For instance, TOF values are very small and approximates to 0.0000 (at 4 decimal places). Feature selection was achieved using correlation matrix (Fig.1). The range of values in the correlation matrix is between ± 1; $+1$ indicates strongly related terms while -1 implies that the variables under consideration has zero relationship. It is best practice to employ features with less or zero correlation in the construction of models.

A simple correlation matrix was done in excel, upon execution, we see in Fig.1 that the rows for Charge and ionMass is populated with the value $DIV/0!$ across all feature columns except their respective columns. This is because 'charge' and 'ionMass' values is '1' for every ion in all assay results. Similarly, the sum of substance Mass and Charge is molecular mass of ions i.e. $m_i = S_i + C_i$.

The highlighted (M/Z - TOFIntensity and SubstanceMass - TOFIntensity) cells indicates the features that were chosen or less correlated since we ignored the TOF due to its size. Sequel to these, each matrix (R_k^S) was reduced to an n-by-2 matrix (P_k^S).

$$P_1^C = \begin{bmatrix} m_1^1 & I_1^1 \\ m_2^1 & I_2^1 \\ \vdots & \vdots \\ m_n^1 & I_n^1 \end{bmatrix} \qquad P_k^M = \begin{bmatrix} m_1^k & I_1^k \\ m_2^k & I_2^k \\ \vdots & \vdots \\ m_n^k & I_n^k \end{bmatrix} \tag{2}$$

The matrix P is basically a collection of peaks with only the m/z (m_n^k) and TOFIntensity (I_n^k) features. Going forward, we shall simply refer to P as a *data point*, m/z as *mass*, TOFIntensity as *intensity* and an ion as a *peak* defined by the pair $(mass, intensity)$.

3.2 Problem Formulation

The problem formulation is based on the hypothesis that the average intra class distance between Mass Spectra ions is significantly less than inter class distances. Let MSpool be the pool of all the data points in the data set, expressed as

$$MSpool = \{(P_1^C), (P_1^M), (P_1^T), \cdots, (P_k^C), (P_k^M), (P_k^T)\}$$

where the superscripts C, M, and T represents the 3 disease stages. The problem is targeted at finding a *selection* $f(m, I)$ such that

$$f(m, I) = f_1(m_1, I_1), f_2(m_2, I_2), \cdots, f_n(m_n, I_n) \tag{3}$$

correctly predicts elements of a particular stage label.

Definition 3.1. Jackknife Procedure: Given a sample (X) of size N, a *delete-d* Jackknife sample is obtained by selecting and deleting 'd' number of observations from the sample. For instance, a delete-1 Jackknife sample

[1] m_n is m/z (or molecular mass), T_n stands for time-of-flight (TOF), I_n denotes TOFIntensity, S_n is Substance mass, C_n for ion charge, M for ion mass, N for signal-to-noise and P implies peak type of the ion

will look like;

$$X_a = X_b, X_c, \cdots, X_n \tag{4}$$

In this case, X_a is the deleted observation, and used as the test data here, while other terms of the equation constitutes the train data-set.

3.3 Exponential Euclidean Distance

In general, there exists three cases that may exist between any two peaks, irrespective of the stage or the data point they belong to. These are 1) equal molecular mass values but different intensity values, 2) unequal molecular mass and intensity values, and 3) unequal mass values but equal intensity values. It is most profitable to align and work with the elements of the data set with the scenario described by case 1. Consequently, it is crucial to scale and quantify the terms $\Delta I_i = I_1 - I_2$ and $\Delta m_i = m_1 - m_2$ differently.

We now introduce equation 5 known as *exponential Euclidean* distance (EED) function, it is biased and an improvement of the Euclidean distance formula. It exponentially blows up the contribution of Δm_i, thereby, filtering out vectors whose combination violates case 1. EED defines the distance between two vectors a and b by

$$dist_{(a,b)} = \sqrt{(e^{(m_a - m_b)^2} - 1)^2 + (I_a - I_b)^2} \tag{5}$$

where m and I respectively represent the mass and intensity values of the row vectors a and b. It is crucial to establish that equation 5 satisfies the norm axioms before it is used. Notice that, it is easy to verify the positivity and symmetricity laws of the norm axioms, hence, we dwell on only the triangular inequality law.

Proof. TRIANGULAR INEQUALITY
Let $dist_{(a,c)}$ be the distance between vectors a and c, while $dist_{(a,b)}$ and $dist_{(b,c)}$ are respectively the distance measures from a to b and b to c. By taking absolute values we know that

$$-|dist_{(a,b)}| \leq dist_{(a,b)} \leq |dist_{(a,b)}| \; ; -|dist_{(b,c)}| \leq dist_{(b,c)} \leq |dist_{(b,c)}| \; ; -|dist_{(a,c)}| \leq dist_{(a,c)} \leq |dist_{(a,c)}| \tag{6}$$

Summing up any two elements of equations 6, for instance, the first two equations yields

$$-(|dist_{(a,b)}| - |dist_{(b,c)}|) \leq dist_{(a,b)} + dist_{(b,c)} \leq |dist_{(a,b)}| + |dist_{(b,c)}| \tag{7}$$

$$Rewritten,\ we\ have, \qquad |dist_{(a,b)} + dist_{(b,c)}| \leq |dist_{(a,b)}| + |dist_{(b,c)}| \tag{8}$$

In equation 5, m_a and I_a are positive terms while m_b and I_b are negative terms. In other words, first variables in each term has positive coefficients while the second variables are with negative (-1) coefficients.

$$Define \qquad X_{m_{ab}} = (e^{(m_a - m_b)^2} - 1)^2; \qquad X_{I_{ab}} = (I_a - I_b)^2 \; ; \tag{9}$$

$$We\ \ have \ \ \sqrt{X_{m_a}} = e^{(m_a)^2} - 1 \; ; \qquad \sqrt{X_{I_a}} = I_a \; , \qquad if\ (m_b, I_b) = 0 \tag{10}$$

$$So \; , \qquad dist_{(a,b)} = \sqrt{X_{m_{ab}} + X_{I_{ab}}} \tag{11}$$

$$\implies (dist_{(a,b)})^2 = X_{m_{ab}} + X_{I_{ab}} = \left(X_{m_a} - X_{m_b}\right) + \left(X_{I_a} - X_{I_b}\right) \tag{12}$$

$$Similarly \; ; \qquad (dist_{(b,c)})^2 = X_{m_{bc}} + X_{I_{bc}} = \left(X_{m_b} - X_{m_c}\right) + \left(X_{I_b} - X_{I_c}\right) \tag{13}$$

Summing elements of equations 11 and 12 and cancelling out like terms yields

$$
\begin{aligned}
(dist_{(a,b)})^2 + (dist_{(b,c)})^2 &= (X_{m_{ab}} + X_{I_{ab}}) + (X_{m_{bc}} + X_{I_{bc}}) && (14)\\
&= \left(X_{m_a} - X_{m_b} + X_{I_a} - X_{I_b}\right) + \left(X_{m_b} - X_{m_c} + X_{I_b} - X_{I_c}\right)\\
&= \left((X_{m_a} - X_{m_c})\right) + \left((X_{I_a} - X_{I_c})\right)\\
&= (dist_{(a,c)})^2 && (15)\\
\implies dist_{(a,c)} &= dist_{(a,b)} + dist_{(b,c)} && (16)
\end{aligned}
$$

Using equations 8 and 16 we have

$$
\begin{aligned}
|dist_{(a,c)}| \leq |dist_{(a,b)} + dist_{(b,c)}| &\leq |dist_{(a,b)}| + |dist_{(b,c)}|\\
\implies |dist_{(a,c)}| &\leq |dist_{(a,b)}| + |dist_{(b,c)}| && (17)
\end{aligned}
$$

□

3.3.1 KNN Distance Hit Table

First, a test data is obtained using the Jackknife description, call the vectors in the test data test vectors. Predicting a test data entails generating a distance *'Hit Table'* that records by vote, the stage labels of the vectors closest to its test vectors using equation 5 and the principle of KNN. In each iteration, equation 5 is used to determine the distance between all possible pairs between vectors from the test data point and corresponding vectors of the train data points i.e., pair of vectors that satisfies case 1 with the test vectors irrespective of stage. Then, the stage label of vectors within k-minimum distance from each test vector is identified and recorded. Below is an example of a typical hit table,

Hit Table			
	CON	MCI	tAD
TEST1	44	**75**	60
TEST2	49	57	**73**

The column titles CON, MCI, and tAD respectively holds counts of vectors in the train data set across the stages that are closest to vectors of a $TEST$ data. At the end, a test data is classified into the stage with the highest number of minimum hits. For example, $TEST1$ is classified to be MCI while $TEST2$ is tAD based on majority vote.

4 Results and Discussion

Using the Jackknife re-sampling technique, each disease stage produced 20 test data-points. Consequently, we performed sixty KNN classification iterations with $k = 1$ for the 3 proteinChips and energy levels. The classification results in percentages is captured in the table below

Classification Results		
ENERGY LEVEL	LOW	HIGH
CM10	53	80
IMAC30	54	82
Q10	56	85

In this paper, we adopted the principle of KNN and introduced a 2-scale distance function to build an exponential Euclidean distance classifier of mass spectra data. Hitherto, differentially expressed ions in SELDI

data are identified, singled out, and match with and/or against ions contained in public data base or theoretical ions in biomarker discovery or protein identifications. The outcome of this study is novel, in that, it teaches a framework to simultaneously mine pools of ions or spectrum with variant backgrounds (stages). In particular, and taking high energy results, we present a new aproach for diagnosing Alzheimer's disease, based on the characterization of a collection of SELDI Saliva Spectra data.

The data structure was the first problem we had to overcome. The decomposition of the feature matrices to a collection of feature vectors, sequel to a systematic feature selection enabled us to solve the problem in a 2-dimensional space.

This work pinpoints inherent pattern(s) in saliva SELDI data (as well as other protein source sample(s)). In general, the classification results of mining the data set generated with low energy laser bombardment leaves much to be desired when compared to that obtained from high energy laser bombardment data set irrespective of the proteinChip used in the analysis.

Using Saliva SELDI data was also a plus owing to fact that it is less invasive to obtain saliva samples. The presence of several molecular mass values but with different intensity values made this KNN approach possible, in that, we were able to geometrically mark the intensities of similar mass values in space and used their geometric location for discrimination.

5 Conclusion and Future Works

It is worthy to ask if adding additional features into the exponential distance function will improve the result of this work. Similarly, will it improve the obtained result if only ions of a particular peak type or if ions are categorized and used based on their molecular weight or signal to noise ratio? Future works can be directed at answering these questions.

A sensitivity analysis of saliva SELDI data with regards to identifying the best time of the day or activities that could precede obtaining of saliva samples from donors is also a possible issue for future work.

Having raised a standard vis-a-vis the ability to identify and classify stage instances with over 80% accuracy, a future work might be directed to dissect elements of MCI stage with a view of predicting the rate with which they could most likely degenerate to acute Alzheimer's disease or recuperate to a no impairment status overtime.

As a goal for future work, the result of this study can be extended to serve as a tool to monitor AD patients conditions since the disease severity status can easily be determined with the number of 'hit points' in the distance table.

In conclusion, it is crucial to have studies focused on closing the gap between known protein bio-markers and diagnosis of incurable diseases (e.g Dementia). They have the needed potentials to saving and improving the quality of lives, lowering health care costs, with positive impacts for personalized medicine.

References

[1] Alzheimer's Association. http://www.alz.org/facts/overview.asp. [Accessed 3 July 2016].

[2] Jae Won Lee, Jung Bok Lee, Mira Park, Seuck Heun Song. An extensive comparison of recent classification tools applied to microarray data. Comput. Stat. Data Anal. 2005; 48:869-885.

[3] José Crossa, Jorge Franco. Statistical Methods for Classifying genotypes. Euphytica. 2004; 137(1):19-87.

[4] Leping Li, Clarice R. Weinberg, Thomas A. Darden and Lee G. Pedersen. Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameter of GA/KNN method. Bioinformatics. 2001; 17(12):1131-1142.

[5] Min-Ling, Zhi-Hua Zhou. ML-KNN: A lazy learning approach to multi-label learning. Pattern Recogn. Society. 2007; 40(7): 2038-2048.

[6] Wangmeng Zuo, David Zhang, Kuanquan Wang. Bidirectional PCA with Assembled Matrix Distance for Image Recognition. Cybernetics. 2006;36(4):863-872.

[7] Liisa Holm and Chris Sander. Protein Structure Comparison by Alignment of Distance Matrices. J. Mol. Biol. 1993; 233(1): 123-138.

[8] Kilian Q. Weinberger, Lawrence K. Saul. Distance Metric Learning for Large Margin Nearest Neighbor Classification. JMLR. 2009; 10: 2007-244.

[9] Hao Zhang, Alexander C. Berg, Michael Maire, Jitendra Malik. SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition. PROC. CVPR. IEEE. 2006; 2:2126-2136.

[10] Avery I. McIntosh. "The Jackknife Estimation Method". arXiv:1606.00497 [stat.ME]. 2016.

[11] Sandra L. Taylor and Kyoungmi Kim. A Jackknife and Voting Classifier Approach to Feature Selection and Classification.Cancer Inform. 2011; 10: 133-147.

[12] Andrew Keller, Alexey I. Nesvizhskii, Eugene Kolker, and Ruedi Aebersold. Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. Anal. Chem. 2002, 74, 5383-5392.

[13] Issaq H, Veenstra T, Conrads T, Felscow D. The SELDI-TOF MS Approach to Proteomics: Protein Profiling and Biomarker Identification. Biochem. Biophys. Res. Commun. 2002; 292:587-597.

[14] G. P. S. Raghava Bioinformatics and drug discovery. http://bioinformaticsweb.net/drugdiscovery.html [Accessed 10 June, 2015]

[15] Jonathan M. Street and James W. Dear The Application of Mass Spectrometry Based Protein Biomaker Discovery to Theragnostics. Br. J. Clin. Pharmacol. 2010; 69(4):367–378

Figure 1: **Correlation Matrix**

	M/Z	*TOF*	*TOFIntensity*	*SubstanceMass*	*Charge*	*IonMass*	*SignalToNoise*	*PeakType*
M/Z	1							
TOF	0.999424	1						
TOFIntensity	-0.05225	-0.0748826	1					
SubstanceMass	1	0.99942415	-0.052247116	1				
Charge	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	1			
IonMass	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	#DIV/0!	1		
SignalToNoise	-0.20222	-0.2241154	0.986469634	-0.20222217	#DIV/0!	#DIV/0!	1	
PeakType	0.037922	0.06308207	-0.774132934	0.037921537	#DIV/0!	#DIV/0!	-0.739889925	1

Parameter Estimation of Stochastic Models Based on Limited Data

Minghan Chen
Virginia Tech
Blacksburg, Virginia
cmhshirl@vt.edu

Yang Cao
Virginia Tech
Blacksburg, Virginia
ycao@cs.vt.edu

Layne T. Watson
Virginia Tech
Blacksburg, Virginia
ltw@cs.vt.edu

ABSTRACT

Progress in experimental techniques enables a more accurate quantification of genes, mRNA, and proteins at the single cell level. Provided with limited time series data from single-cell measurements, this note proposes a new quasi-Newton optimization algorithm (QNSTOP) for parameter estimation of stochastic models. To capture the stochasticity inside models and data, the random objective function is constructed based on the maximum log-likelihood of transition probabilities rather than summary statistics, which relies heavily on stochastic simulations. Simple to use and efficient, QNSTOP can find the 'best' parameter vector from far away starting points in just a few iterations. Results on a bistable model match well the bistable dynamics that can only be obtained from stochastic models.

CCS CONCEPTS

• **Mathematics of computing → Mathematical optimization**;
• **Applied computing → Systems biology**;

KEYWORDS

stochastic models, parameter estimation, QNSTOP

ACM Reference Format:
Minghan Chen, Yang Cao, and Layne T. Watson. 2017. Parameter Estimation of Stochastic Models Based on Limited Data. In *Proceedings of 2017 (SIGBio Newsletter).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

With the improvement in experimental techniques [13], biologists are able to quantify genes and proteins and their dynamics in a single cell. These experimental data call for quantitative stochastic models for gene and protein networks at cellular levels that match well with the data and account for cellular noise. Stochastic methods, such as Gillespie's Stochastic Simulation Algorithm (SSA) [6], can be used to calculate the probability solution of the chemical master equation (CME) [7]. The CME generally results in a linear system of very high or infinite dimension, which in the latter case is unsolvable. Many methods have been proposed to approximate the solution of the CME, such as moment closure methods [8], radial basis functions [15], and the finite state projection (FSP) method [11].

Assuming that a chemical reaction network is developed, an important and difficult part of modeling is the estimation of model parameters. For stochastic models, parameter estimation is even more challenging as the amount of empirical data must be large enough to obtain statistically valid parameter estimates. Various approaches have been developed to estimate parameters for stochastic biochemical systems, from Bayesian statistics [12, 14], to control theory [9, 10], and to optimization techniques [2, 12].

This note presents a quasi-Newton algorithm (QNSTOP) [1] for stochastic optimization problems, which worked well for a state-of-the-art stochastic model of the budding yeast cell cycle containing 52 unknown parameters [3]. Different from matching ensemble statistics, here this algorithm is applied to a set of limited data, such as a single trajectory (observation) of the stochastic process, or even just part of the data.

2 MAXIMUM LOG-LIKELIHOOD

To capture the stochastic fluctuations, measure the transition probability that a system jumps from one state to the next state after a certain time step. Suppose $D = [x_1, x_2, ..., x_m]$ is a sequence of the molecule numbers of a certain species in a cell collected after every time period τ, where m is the data size. The logarithm of the likelihood of the observed data D is

$$\log \mathcal{L}(\theta|D) = \log\left(\prod_{k=1}^{m-1} E_{x_k, x_{k+1}}\right) = \sum_{k=1}^{m-1} \log E_{x_k, x_{k+1}} \tag{1}$$

where $\theta \in \mathbb{R}^n$ are model parameters and E is the transition matrix. Specifically, $E_{i,j}$ is the transition probability that the system changes from state i to state j. A larger value of log-likelihood indicates a higher similarity between the empirical data and simulation data with parameter vector θ.

The objective function is

$$f(\theta) = -\log \mathcal{L}(\theta|D), \tag{2}$$

and the stochastic optimization problem to be solved is

$$\min_{\theta \in \Theta} f(\theta), \tag{3}$$

where Θ is a set in $\mathbb{R}^n$ defining the feasible set (allowable values for the model parameter vector θ).

3 A BISTABLE MODEL

To simplify the presentation, consider a bistable model consisting of only one species S, which has a positive feedback on itself (shown in Figure 1) This single species positive feedback model demonstrates a stochastic switching behavior that does not fit deterministic models.

In general, consider a well-mixed system of N distinct species and M reaction channels. The chemical master equation is

$$\frac{\partial}{\partial t} P(X;t) = P(X;t)A, \tag{4}$$

SIGBio Newsletter, 2017
2017. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

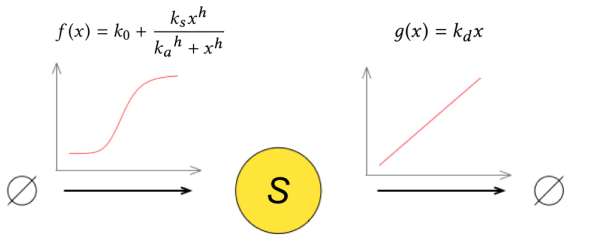

Figure 1: A simple bistable model with $f(x)$ as the synthesis function and $g(x)$ as the degradation function. There are five parameters in the model: the degradation rate k_d; the base rate for the Hill equation k_0, the rate constant for synthesis k_s, the dissociation constant k_a, and the Hill coefficient h.

where $X = [x_1, x_2, \ldots]$ is all possible population vectors x_i at time t and $P(X;t)$ represents the probabilities of those population vectors at time t. A is the state reaction matrix

From equation (4), the transition probability matrix can be calculated by

$$E = e^{A\tau}, \tag{5}$$

where τ is the amount of time the system has evolved from a previous time. A is the state reaction matrix, written as:

$$A_{ij} = \begin{cases} -\sum_{\mu=1}^{M} a_\mu(x_j), & \text{for } i = j, \\ a_\mu(x_i), & \text{for } i \text{ such that } x_j = x_i + v_\mu, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

where v_μ is the stoichiometric transition vector for channel μ.

When a system has an infinite number of states, such as the aforementioned bistable model, the finite state projection (FSP) method [11] projects the infinite state vector X to a finite state vector, approximating the CME solution with an error ϵ. Accordingly, A, E are approximated by $\hat{A}, \hat{E}$. Fox et al. [5] proved that the FSP-derived likelihood converges monotonically to the exact likelihood value.

4 QUASI-NEWTON ALGORITHM FOR STOCHASTIC OPTIMIZATION

QNSTOP is a class of quasi-Newton methods developed for optimization of a stochastic objective function $f(X)$ (X is the parameter vector). A variant of QNSTOP has also been quite successful for deterministic global optimization problems [4]. In the k-th iteration, QNSTOP computes the gradient vector $\hat{g}_k$ and the Hessian matrix $\hat{H}_k$ of a quadratic model

$$\widehat{m}_k(X - X_k) = \hat{f}_k + \hat{g}_k^T (X - X_k) + \frac{1}{2} (X - X_k)^T \hat{H}_k (X - X_k)$$

of the objective function f centered at X_k, where $\hat{f}_k$ is generally not $f(X_k)$. The next iterate is

$$X_{k+1} = \left(X_k - \left[\hat{H}_k + \mu_k W_k \right]^{-1} \hat{g}_k \right)_\Theta,$$

where μ_k is the Lagrange multiplier of a trust region subproblem, W_k is a symmetric, positive definite scaling matrix, and $(\cdot)_\Theta$ denotes projection onto the feasible set Θ.

To estimate the gradient, QNSTOP uses an ellipsoidal design region of radius τ_k centered at the current iterate $X_k \in \mathbb{R}^n$, given by

$$E_k(\tau_k) = \left\{ X \in \mathbb{R}^n : (X - X_k)^T W_k (X - X_k) \le \tau_k^2 \right\},$$

where W_k is a scaling matrix in

$$W_\gamma = \left\{ W \in \mathbb{R}^{n \times n} : W = W^T, \ \det(W) = 1, \ \gamma^{-1} I_n \preceq W \preceq \gamma I_n \right\}$$

for some $\gamma \ge 1$, where I_n is the $n \times n$ identity matrix. (The notation $A \preceq B$ means that $B - A$ is positive definite.)

In each iteration, QNSTOP chooses a set of N uniformly sampled design sites $\{X_{k1}, \ldots, X_{kN}\} \subset E_k(\tau_k) \cap \Theta$. Let $Y_k = (y_{k1}, \ldots, y_{kN})^T$ denote the N-vector of responses modeled by the linear model $y_{ki} = \hat{f}_k + X_{ki}^T \hat{g}_k + \epsilon_{ki}$, where ϵ_{ki} accounts for lack of fit. $\hat{g}_k$ is then the least squares estimate of the linear model gradient.

For numerical robustness, QNSTOP constrains the Hessian matrix update to satisfy

$$-\eta I_n \preceq \hat{H}_k - \hat{H}_{k-1} \preceq \eta I_n$$

for some $\eta \ge 0$, using a variation of the SR1 (symmetric, rank one) quasi-Newton update described in detail in [1].

QNSTOP utilizes an ellipsoidal trust region concentric with the design region for controlling step length. In one option to the code (mode 'G'), the trust region ellipsoid radius ρ_k is taken equal to the design ellipsoid radius τ_k, and the optimization problem

$$\min_{X \in E_k(\rho_k)} \hat{g}_k^T (X - X_k) + \frac{1}{2} (X - X_k)^T \hat{H}_k (X - X_k)$$

is solved for X_{k+1} and μ_k related by

$$X_{k+1} = X(\mu_k) = X_k - \left[\hat{H}_k + \mu_k W_k \right]^{-1} \hat{g}_k.$$

In another option to the code (mode 'S'), μ_{k-1} is directly updated to μ_k, giving $X_{k+1} = X(\mu_k)$ as above. If necessary, X_{k+1} is projected back into the feasible set Θ.

Finally, the experimental design region $E_k(\tau_k)$ is updated to approximate a confidence set by updating the scaling matrix W_k. The updated scaling matrix is given by

$$W_{k+1} = \left(\hat{H}_k + \mu_k W_k \right)^T V_k^{-1} \left(\hat{H}_k + \mu_k W_k \right),$$

where V_k is the covariance matrix of $\nabla \widehat{m}_k(X_{k+1} - X_k)$.

For numerical stability, W_{k+1} is constrained (by modifying its eigenvalues) to satisfy the constraints $\gamma^{-1} I_n \preceq W_{k+1} \preceq \gamma I_n$ and $\det(W_{k+1}) = 1$, so $W_\gamma \ni W_{k+1}$.

5 NUMERICAL RESULTS

The empirical data fundamentally determines the quality of estimated parameters regardless of the optimization algorithms used. Take the time window of data collection as an example, large time steps τ may miss details of fast reactions, and small time steps in short windows may lose the information from slow reactions. Empirical data was generated by using the exact values in Table 1

as the ground "truth" and collecting $m = 100$ data points every two time units ($\tau = 2$) from the stable state in one SSA simulation.

First consider finding the two parameters k_s and k_d with the rest fixed. In Figure 2, starting from the upper bound of a four-times larger search box $[L/4, 4U]$, QNSTOP quickly finds the exact values of k_s and k_d and settles down. Applying QNSTOP to the full parameter vector, the objective function drops sharply in the first few iterations and oscillates around the minimum value (≈ 600). Note that the objective function at the ellipsoid center and at the best sampled point inside that ellipsoid are different in the beginning, indicating the variability of the stochastic objective function values within that ellipsoid. As QNSTOP iterates, the design ellipsoid radius decreases, and so does the variability. This suggests that the inherent simulation variance for a fixed parameter vector is comparable to the variance within the (small) design ellipsoid.

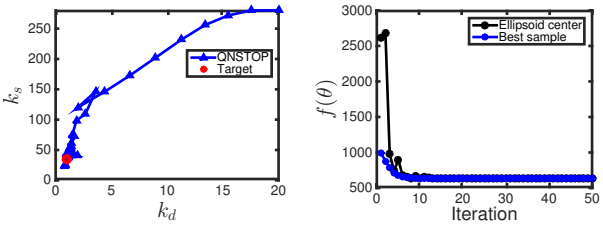

Figure 2: Left: QNSTOP traces of parameters k_s and k_d starting from the upper bound $4U$. Right: execution traces of QNSTOP with L as the initial values.

Table 1 lists the best parameter vector found by QNSTOP that has the minimum objective function value in the search box $[L, U]$. Most parameters are close to the exact value, except for k_0 and k_s. The final distribution of species S population using the best value fits well with that of the exact value in Figure 3, indicating that the two parameters k_0 and k_s are less sensitive in the bistable model.

Table 1: List of parameters in the bistable model.

Parameter	Best value	Exact value	$[L, U]$
k_0	15.22	10.0	[0.1, 20.0]
k_s	50.40	35.0	[0.1, 70.0]
k_a	25.87	25.0	[0.1, 50.0]
h	6.31	6	[0.1, 12.0]
k_d	1.42	1.0	[0.1, 5.0]

6 CONCLUSIONS

With limited time course data, QNSTOP can quickly find parameters that produce similar model dynamics that only exist in stochastic models. The method proposed is simple to use and efficient for models involving fast reactions, which are time consuming for simulation based methods. For large and complex systems where it is hard to compute the transition matrix, QNSTOP is still applicable by using a different stochastic objective function. In fact, QNSTOP has been applied to such models and works well matching empirical data [3].

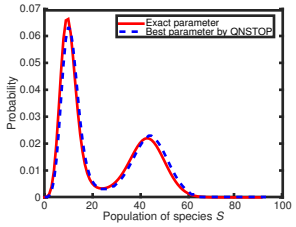

Figure 3: Population distributions of species S using exact parameter and best value found by QNSTOP.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation under award CCF-1526666 and MCB-1613741, and by the 2016 Q-bio Summer School at Fort Collins, CO.

REFERENCES

[1] Brandon D. Amos, David R. Easterling, Layne T. Watson, William I. Thacker, Brent S. Castle, and Micheal W. Trosset. 2014. *Algorithm XXX: QNSTOP—quasi-Newton algorithm for stochastic optimization.* Dept. of Computer Sci. TR-142. Virginia Tech, Blacksburg, VA.

[2] Maksat Ashyraliyev, Johannes Jaeger, and Joke G. Blom. 2008. Parameter estimation and determinability analysis applied to Drosophila gap gene circuits. *BMC Systems Biology* 2 (2008), 83. https://doi.org/10.1186/1752-0509-2-83

[3] Minghan Chen, Brandon D. Amos, Layne T. Watson, John J. Tyson, Yang Cao, Clifford A. Shaffer, Michael W. Trosset, Cihan Oguz, and Gisella Kakoti. 2017. Quasi-Newton stochastic optimization algorithm for parameter estimation of a stochastic model of the budding yeast cell cycle. (2017). To be published.

[4] David R. Easterling, Layne T. Watson, Michael L. Madigan, Brent S. Castle, and Michael W. Trosset. 2014. Parallel deterministic and stochastic global minimization of functions with very many minima. *Computational Optimization and Applications* 57, 2 (01 Mar 2014), 469–492. https://doi.org/10.1007/s10589-013-9592-1

[5] Zachary Fox, Gregor Neuert, and Brian Munsky. 2016. Finite state projection based bounds to compare chemical master equation models using single-cell data. *The Journal of Chemical Physics* 145, 7 (08 2016), 074101.

[6] Daniel T. Gillespie. 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* 22, 4 (Dec. 1976), 403–434. https://doi.org/10.1016/0021-9991(76)90041-3

[7] Daniel T. Gillespie. 2007. Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry* 58, 1 (2007), 35–55. https://doi.org/10.1146/annurev.physchem.58.032806.104637 PMID: 17037977.

[8] Joao Hespanha. 2008. Moment closure for biochemical networks. In *3rd International Symposium on Communications, Control and Signal Processing*. 142–147. https://doi.org/10.1109/ISCCSP.2008.4537208

[9] Niels Rode Kristensen, Henrik Madsen, and Sten Bay Jørgensen. 2004. Parameter estimation in stochastic grey-box models. *Automatica* 40, 2 (2004), 225–237. https://doi.org/10.1016/j.automatica.2003.10.001

[10] Gabriele Lillacci and Mustafa Khammash. 2010. Parameter estimation and model selection in computational biology. *PLOS Computational Biology* 6, 3 (03 2010), 1–17. https://doi.org/10.1371/journal.pcbi.1000696

[11] Brian Munsky and Mustafa Khammash. 2006. The finite state projection algorithm for the solution of the chemical master equation. *The Journal of Chemical Physics* 124, 4 (2006), 044104. https://doi.org/10.1063/1.2145882

[12] Suresh Kumar Poovathingal and Rudiyanto Gunawan. 2010. Global parameter estimation methods for stochastic biochemical systems. *BMC Bioinformatics* 11 (2010), 414. https://doi.org/10.1186/1471-2105-11-414

[13] Arjun Raj and Alexander van Oudenaarden. 2009. Single-molecule approaches to stochastic gene expression. *Annual review of biophysics* 38 (2009), 255–270. https://doi.org/10.1146/annurev.biophys.37.032807.125928

[14] Stefan Reinker, Rachel M. Altman, and Jens Timmer. 2006. Parameter estimation in stochastic biochemical reactions. *IEE Proceedings - Systems Biology* 153, 4 (July 2006), 168–178. https://doi.org/10.1049/ip-syb:20050105

[15] Jingwei Zhang, Layne T. Watson, Christopher A. Beattie, and Yang Cao. 2010. Radial basis function collocation for the chemical master equation. *International Journal of Computational Methods* 07, 03 (2010), 477–498.

ACM BCB 2018 Preliminary announcement
http://acm-bcb.org/2018/index.php

The 9th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM BCB) is the flagship conference of the ACM SIGBio. ACM-BCB 2018 is the conference's ninth year, building upon the success of the first eight meetings in Boston, Niagara Falls, Chicago, Orlando, Washington DC, Newport Beach, Atlanta, and Seattle. ACM-BCB 2018 will be held in Washington, D.C., August 29th -September 1st, 2018.

The conference is the premier dissemination forum for interdisciplinary research linking computer science, mathematics, statistics, biology, bioinformatics, biomedical informatics, and health informatics. The past few decades have seen tremendous growth in the scale and complexity of biological and medical data including recent mainstream recognition of big data challenges. This conference serves to showcase leading-edge research on new technologies and techniques around gathering, processing, analyzing, and modeling of data and information for a variety of scientific, clinical, and healthcare applications, from bench to bedside.

Organizing Committee
General Chairs:
Amarda Shehu, George Mason University
Cathy Wu, University of Delaware

Program Chairs:
Mihai Pop, University of Maryland College Park
Jing Li, Case Western Reserve
Christina Boucher, University of Florida at Gainesville
Liu Hongfang, Mayo Clinic

Web site: http://acm-bcb.org/2018/index.php