Metabolomic Data Profiling to Diabetes Research in Qatar and Middle East Region

Raghvendra Mall
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
Email: rmall@qf.org.qa

Laure Berti
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
Email: lberti@qf.org.qa

Halima Bensmail
Qatar Computing Research Institute
Hamad Bin Khalifa University
Doha, Qatar
Email: hbensmail@qf.org.qa

***Abstract*—Diabetes is a leading health problem in the developed world. The recent surge of wealth in Qatar has made it one of the most vulnerable nations to diabetes and related diseases. Recent technological advances in 1H nuclear magnetic resonance (NMR) spectroscopy techniques for metabolomics profiling offer a great opportunity for biomarkers discovery. Using this technology, we present in this study, an integrative approach with a recently proposed algorithm [1] named Kernel Spectral Clustering to discover new metabolites and possibly new biomarkers. We performed an integrative analysis of 1H NMR spectras measured in urine, from 348 participants of the Qatar Metabolomics Study on Diabetes (QMDiab). Our analyses revealed grouped metabolites that correlate with diabetes and identified specific metabolites affected by antidiabetes medication, which constraints differentiation between diabetic and control patients.**

I. Introduction

Many chronic diseases like Type II Diabetes (T2D) and its complications may be preventable by avoiding factors that trigger the disease process. Accurate prediction and identification using biomarkers will be useful for disease prevention and initiation of proactive therapies to those individuals who are most likely to develop the disease. Recent technological advances in proton 1H Nuclear Magnetic Resonance (NMR) spectroscopy techniques for metabolomics profiling offer great opportunity for biomarker discovery ([8]-[6]). Because of experimental issues in the technical equipment, the levels of some metabolites cannot be universally determined. As the number of measured metabolites often exceeds the number of samples, dimensionality reduction methods are required. Abundances of metabolites are indicative of a variety of conditions, and can provide important insights in a wide variety of biological and clinical investigations. At the same time, interpretation of the spectra gives rise to substantial methodological challenges. The spectra are subject to biological and technical variations, and to uncertainty in identification and quantification of peaks. Nuclear magnetic resonance spectroscopy is a method of choice for identifying and quantifying metabolites in complex biological mixtures, as it is fast, non-destructive and highly reproducible. However interpretation of the spectra is hampered by their complexity, presence of overlapping peaks, and biological variation in the abundance of metabolites. The difficulty is particularly apparent in modern investigations, which require an accurate and fast analysis of spectra from hundreds and even thousands of biological samples. Statistical inference is the only approach that can yield objective and reproducible conclusions from such data. At present the statistical tools available for this task are of limited performance.. We need thorough statistical filtering and biological knowledge integration to handle data noise in every stage of the omics cascade. Technologies to measure high-throughput biomedical data in proteomics, chemometrics, and genomics have led to a proliferation of high-dimensional data that pose many statistical challenges. As metabolites, are biologically interconnected, the variables, in these data sets are not only far larger than the sample size but are often highly correlated and noisy. More generally, methods such as PLS, PCA and SPCA can be used as dimension reduction techniques that finds projections of the data that maximize the covariance between the data and the response [5]. During the last decade, several work have been proposed to encourage sparsity in these projections, or loadings vectors, to select relevant features in high-dimensional data . There are several motivations for regularizing the PCA loadings vectors. Several authors have shown that the PCA projection vectors are asymptotically inconsistent in high-dimensional settings and encouraging sparsity in the loadings has been shown to yield consistent projections [4] However, the computational cost is expensive when requiring a large number of loading so it is desirable to find an approach, which regularize loading scores, reduce features and boost the computation of PCA. The PCA loading vectors can be used as a data compression technique when making future predictions; sparsity further compresses the data. As many variables in high-dimensional data are noisy and irrelevant, sparsity presents a method

for automatic feature selection. This leads to results that are easier to interpret and visualize. While sparsity in PCA is important for high-dimensional data, there is also a need for more general and flexible regularized methods. Consider our NMR spectroscopy as a motivating example. This high-throughput data measures the spectrum of chemical resonances of all the latent metabolites, or small molecules, present in a biological sample. Typical experimental data consists of discretized, functional, and nonnegative spectra with variables measuring in the thousands for only a small number of samples. Additionally, variables in the spectra have complex dependencies arising from correlation at adjacent chemical shifts, metabolites resonating at more than one chemical shift, and overlapping resonances of latent metabolites. Because of these complex dependencies, there is a long history of using PCA to reduce the NMR spectrum for supervised data [3] or sparse PCA [8]. Classical PCA or Sparse PCA, however, are not optimal for this type of data as they do not account for the nonnegativity or functional nature of the spectra and do not encourage sparsity or group sparsity. In this paper, we seek a more flexible and fast framework for analyzing high-dimensional ^{1}H NMR data that encourage sparsity, group sparsity, or smoothness, and also leads to a more computationally efficient and fast numerical algorithm.

II. Metabolic Data Genertion and binning

This study was embedded in the Qatar Metabolomics Study on Diabetes (QMDiab), a cross-sectional case-control study with 348 subjects. The work was a joint collaboration between Hamad Medical Corporation and Weill Cornell Medical College Qatar. Patients were asked to enroll between February and June 2012. The study measured metabolites in 348 individuals within the age of 17 to 81. The metabolites were measured in the three body fluids non-fasting blood plasma, urine, and saliva. In the time from February to June 2012, 1107 samples were taken from the participants, comprising 1563 metabolites including amino acids, peptides, carbohydrates and lipids, as well as age, gender, ethnicity, weight, height, Body Mass Index (BMI) and personal history of T2D. More details is given in [8]. This data comprises 200 Arabs, 99 south Asians, 35 Filipino and 14 other nationalities. Measurement was also made on 173 females and 175 males.
When dealing with high resolution NMR spectra it is in general impracticable to work with the entire data points of the spectra which are usually in the order of 32Kb and bigger. The most common strategy used to reduce the number of variables consists in dividing each spectrum in a defined number of regions, the so called bins. Several binning strategies are available today, from regular binning, where bins have fixed width, to more sophisticated strategies such as gaussian or dynamic adaptive binning [2].
The proposed idea is that in order to handle noisy overlapping data we first perform denoising using the Savitzky-Golay filter [9] approach to KPCA. We obtain the optimal model parameters σ using the Model selection criterion based on Distance Distributions (MDD) as in [10] . After denoising, we cluster the data using the same model parameter σ as obtained from MDD along with user provided number of clusters k to obtain groups with better cluster generalizations

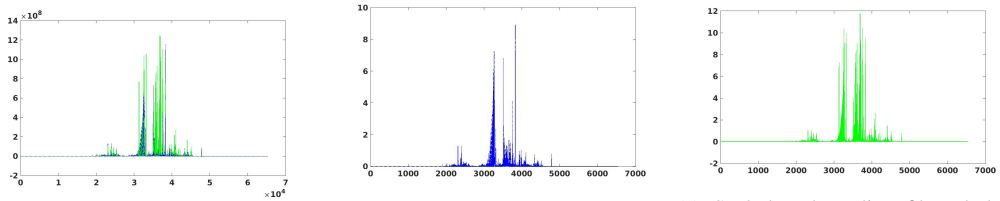
III. Pre-Processing

We first perform an intensive pre-processing of the raw metabolite dataset in order to extract meaningful patterns from the dataset. The raw metabolite dataset comprises of NMR signal information for each sample where this signal information is often very noisy. The first step that we take is to smoothen the signal information corresponding for each sample using the Savitzky-Golay filter [9]. This filter helps to track the signal closely and accounts for the transient effects at the beginning and end of the signal. It performs polynomial fitting to frames of data. They are particularly effective for noisy data, preserving the high frequency components of data while smoothing it. For our experiments we ue the third order polynomial along with frame size $= 25$ for smoothing the original raw metabolite sample signal. Figure 1a gives the signal representation for the entire dataset after smoothing.

After smoothing each sample, we divide the signal information into equi-width bins of size $= 10$. We then use the median filter i.e. we represent the signal value for each bin as the median of the values corresponding to the 10 measurements that comprise a bin. This helps to reduce the part of the signal which is not active (0 values) and capture the informative measurements in the signal while preventing it to be effected by outliers. By undertaking this step we reduce the length of the original signal by a factor of 10. We then divide the dataset into the patients and controls. We undertake a scaling step to scale the values for each bin of each sample in patients and controls respectively by dividing the measurement by a factor of 10^8 as depicted in Figure 1b, 1c. Finally, we remove those bins corresponding to which the signal values are ≈ 0 for more than half the control cases. We take a similar step for the patient dataset thereby reducing the length of active bins.

IV. Kernel Spectral Clustering

We used the Kernel Spectral Clustering (KSC) proposed in [11] for clustering the metabolite dataset. We also used a recently proposed denoised KSC [12] where we first denoise the metabolite dataset using kernel principal component analysis (PCA) with



(a) Original data after smoothing.

(b) Scaled and median filered data for control cases.

(c) Scaled and median filtered data for patient cases.

Fig. 1: Signal representation of the data

model selection based on distance distributions [10]. We then use the KSC method on this noise-free version of the dataset. Below we provide a brief summary of the KSC technique.

A. Primal-Dual Weighted Kernel PCA framework

Given a dataset $\mathcal{D} = \{x_i\}_{i=1}^N$, $x_i \in \mathbb{R}^d$ and the number of clusters k, the primal problem of the spectral clustering via weighted kernel PCA is formulated as follows [1]:

$$\begin{aligned} \min_{w^{(l)},e^{(l)},b_l} \quad & \frac{1}{2}\sum_{l=1}^{k-1} w^{(l)\intercal} w^{(l)} - \frac{1}{2N}\sum_{l=1}^{k-1} \gamma_l e^{(l)\intercal} D_\Omega^{-1} e^{(l)} \\ \text{such that} \quad & e^{(l)} = \Phi w^{(l)} + b_l 1_N, l = 1,\ldots,k-1, \end{aligned} \tag{1}$$

where $e^{(l)} = [e_1^{(l)},\ldots,e_N^{(l)}]^\intercal$ are the projections onto the eigenspace, $l = 1,\ldots,k-1$ indicates the number of score variables required to encode the k clusters, $D_\Omega^{-1} \in \mathbb{R}^{N\times N}$ is the inverse of the degree matrix associated to the kernel matrix Ω. Φ is the $N \times d_h$ feature matrix, $\Phi = [\phi(x_1)^\intercal;\ldots;\phi(x_N)^\intercal]$ and $\gamma_l \in \mathbb{R}^+$ are the regularization constants. The kernel matrix Ω is obtained by calculating the similarity between each pair of data points in the training set. Each element of Ω, denoted as $\Omega_{ij} = K(x_i,x_j) = \phi(x_i)^\intercal\phi(x_j)$ is obtained by using the RBF kernel. The clustering model is given as:

$$e_i^{(l)} = w^{(l)\intercal}\phi(x_i) + b_l, i = 1,\ldots,N, \tag{2}$$

where $\phi : \mathbb{R}^d \to \mathbb{R}^{d_h}$ is the mapping to a high-dimensional feature space d_h, b_l are the bias terms, $l = 1,\ldots,k-1$. The projections $e_i^{(l)}$ represent the latent variables of a set of $k-1$ binary cluster indicators given by $\text{sign}(e_i^{(l)})$ which can be combined with the final groups using an encoding/decoding scheme. The decoding consists of comparing the binarized projections w.r.t. codewords in the codebook and assigning cluster membership based on minimal Hamming distance. The dual problem corresponding to this primal formulation is:

$$D_\Omega^{-1} M_D \Omega \alpha^{(l)} = \lambda_l \alpha^{(l)}, \tag{3}$$

where M_D is the centering matrix which is defined as $M_D = \text{I}_N - (\frac{(1_N 1_N^\intercal D_\Omega^{-1})}{1_N^\intercal D_\Omega^{-1} 1_N})$. The $\alpha^{(l)}$ are the dual variables and the positive definite kernel function $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ plays the role of similarity function. This dual problem is closely related to the random walk model as shown in [1]. The KSC formulation is just a weighted version of KPCA where the weighting term is D_Ω^{-1}.

V. Result

Binning, smoothing, filtering were used to preprocess data. Then kernel spectral clustering, was applied to examine similarities and/or differences in the ^{1}H-NMR spectra. Figure 2 shows kernel matrix or the similarity of the processed urine samples with uniform 0.007 ppm bin widths for patients and control. As we can see clearly, there are two major group of metabolites (yellow and blue) expressed for patients and controls. For patients, the number of expressed metabolites is larger than the number of the ones expressed for control group.
The most important discovery is from Figure 3. We can see that for Control, three groups of metabolites were defined (red, black and bleu) that are associated with three groups of people.

References

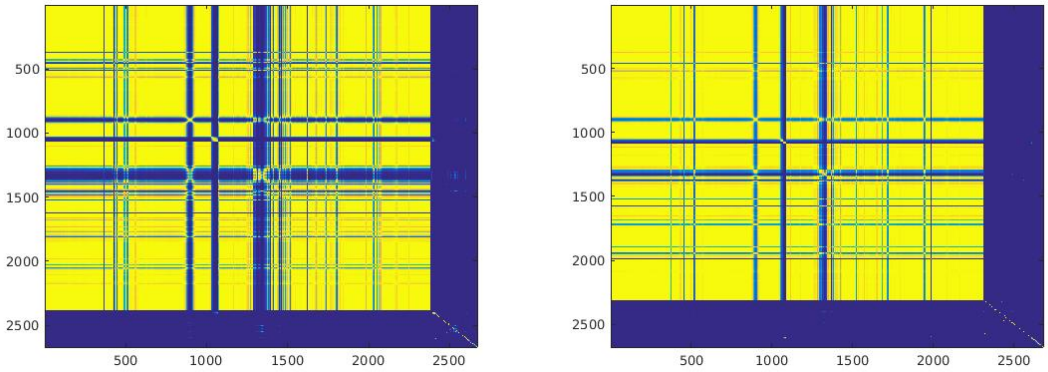
[1] Alzate, C. and Suykens, J.A.K. Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, *32(2)*, pp. 335-347 .

[2] Hackstadt A. J. (2009) Filtering for increased power for microarray data analysis. BMC Bioinformatics, 10:11.

[3] Goodacre R, Vaidyanathan S, Dunn WB, Harrigan GG, Kell DB (2004) Metabolomics by numbers: acquiring and understanding global metabolite data.Trends Biotechnol 22: 245-252.

[4] Zou H., Hastie T., and Tibshirani R. (2006) Sparse Principal Component Analysis.We present a new approach to principal component analysis, that allows us to use an L1 penalty to ensure sparseness of the loadings. JCGS, 15(2): 262-286.

[5] Genevera I. Allen G. I., Christine Peterson C., Vannucci M., and Maletic-Savatic M. (2013) Regularized Partial Least Squares with an Application to NMR Spectroscopy. Stat Anal Data Min. 6(4): 302314.



(a) Similarity measure of processed urine samples with uniform 0.007 ppm bin widths for Control.

(b) Similarity measure of processed urine samples with uniform 0.007 ppm bin widths for Patient.

Fig. 2: Comparision of kernel matrix of metabolite information between control vs patients.

[6] Suhre K, Meisinger C, Doring A, Altmaier E, Belcredi P, Gieger C, Chang D, Milburn MV, Gall WE, Weinberger KMet al. 2010 Metabolic footprint of diabetes: a multiplatform metabolomics study in an epidemiological setting. PLoS ONE

[7] Do K. T. (2013). Metabolomic analysis of multiple bodyfluids in the qatar metabolomics study of diabetes. Master?s thesis, Mnchen, Deutschland: Institute of Computational Biology, Helmholtz Zentrum.

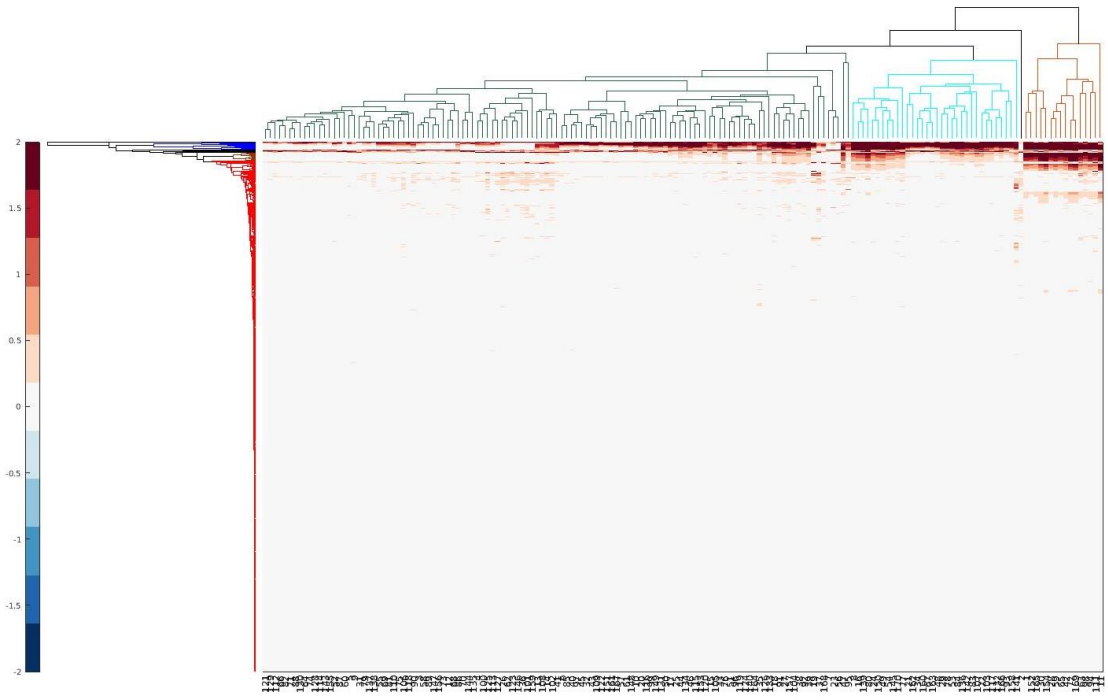
[8] Ullah E., Shahzad M., Rawi R., Dehbi M., Suhre K., Selim M., Mook D., and Bensmail H (2015) Integrative 1H-NMR-based Metabolomic Profiling to Identify Type-2 Diabetes Biomarkers: An Application to a Population of Qatar. Metabolomics 5, 136.

[9] Savitsky, A. and Golay, M.J.E. Smoothing and differentiation of data by simplified least squares procedures, *Analytical Chemistry*, 1971, *36(8)*, 1627-1639.

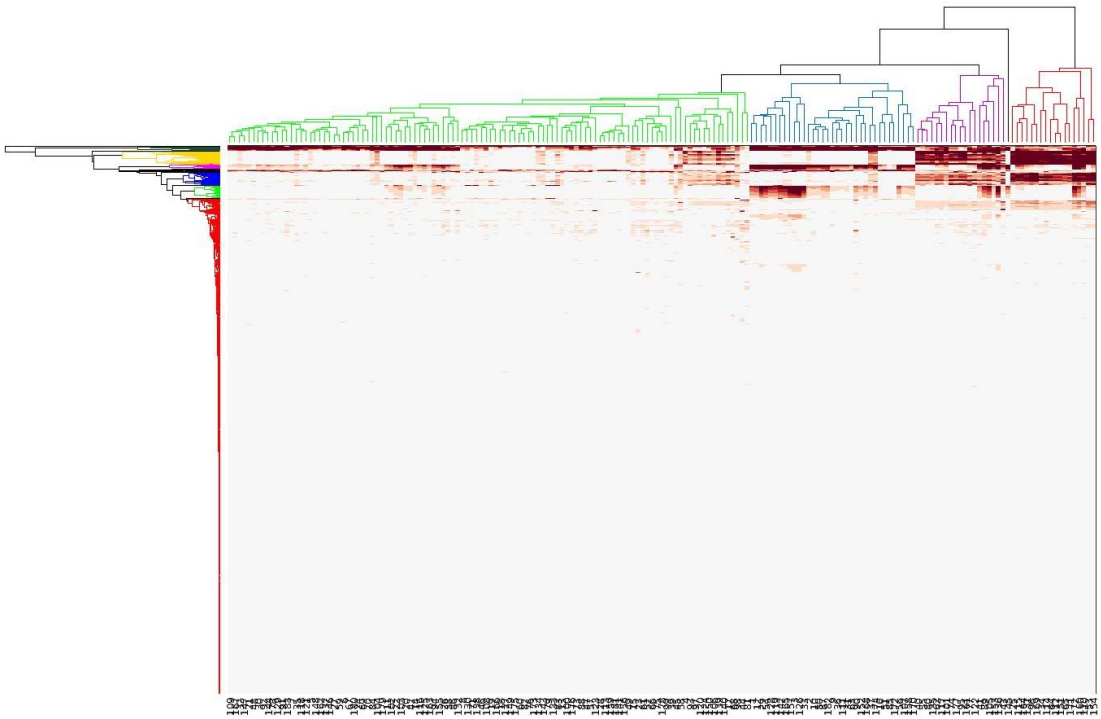
[10] Varon, C., Alzate, C. and Suykens, J.A.K. Noise level estimation for model selection in kernel PCA denoising, *IEEE Transactions on Neural Networks and Learning Systems*, 2015, *26(11)*, 2650-2663.

[11] Mall, R., Langone, R. and Suykens, J.A.K. Kernel Spectral Clustering for Big Data Networks, *Entropy*, 2013, *15(5)*, 1567-1586.

[12] Mall, R., Bensmail, H., Varon, C., Langone, R. and Suykens, J.A.K. Denoised Kernel Spectral Data Clustering, Internal Report, 2016, QCRI.



(a) Biclustering approach for Control.



(b) Biclustering approach for Patient.

Fig. 3: Hierarchical biclustering of samples and metabolite information for patients and controls.