Sparse Network-regularized Nonnegative Matrix Factorization and Applications to Tumor Subtyping

By

Xue Zhong

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

in

Biostatistics

August, 2015

Nashville, Tennessee

Approved:

Yu Shyr, PhD

Xi Chen, PhD

To my amazing sons, Nathan and Nolan, cooperative most of the time

and

To my beloved husband, Bingshan, tremendously supportive

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to my advisor Dr. Yu Shyr, for his patience, motivation and immense knowledge, and providing me financial support through my graduate studies. I also want to thank my co-advisor Dr. Xi Chen, for his insightful comments and valuable suggestions on my thesis work.

I am especially indebted to Dr. Jeffrey Blume, our Director of Graduate Studies, who is always willing to help and give his best suggestions, and who has been supportive of my career goals and has provided me professional guidance. I am grateful to the excellent faculty of the Department of Biostatistics. As my teachers, their dedication and insightfulness have made the learning process a journey full of intellectual fun and fulfillment. I want to thank Dr. Frank E. Harrell, Chair of the Department of Biostatistics, for his excellent leadership to have made the department such a great place to learn and grow.

Special thanks goes to all graduate students in Biostatistics with whom I had the pleasure to work or interact, and my colleagues from the Center for Quantitative Sciences. I would also like to thank our program coordinator Linda Wilson, for her excellent support, and Jill Shell, the administrative manager of Center for Quantitative Sciences, for her excellent assistance.

Finally, my deepest gratitude goes to my family, my parents, my in-laws, my husband Bingshan, and my two wonderful children, Nathan and Nolan. They have provided me with unending love, support, encouragement and inspiration.

LIST OF TABLES

LIST OF FIGURES

TABLE OF CONTENTS

Page

CHAPTER I

INTRODUCTION

Non-negative matrix factorization (NMF) is an unsupervised learning method for finding parts-based decompositions. NMF was formally introduced as a method for face image decompositions[1]. In this setting, NMF yielded a decomposition of human faces into parts resembling features such as eyes, noses etc. The constraint of non-negativity is natural for a wide range of natural signals, such as pixel intensities, occurrence counts, amplitude spectra, and gene expressions. NMF has found wide application in many areas, and has for example been used in image processing[1] to find meaningful features in image datasets; in text processing[2] to find sets of words that constitute latent topics in emails; in audio processing[3] to find separate mixtures of audio sources; and in bioinformatics [4] to find biologically meaningful cancer subtypes based on gene expressions.

Variants of NMF have been proposed for different purposes. In image processing, it was noted early that NMF does not always result in parts-based representations when the face images are not well aligned or in presence of lighting variations[5]; sparse NMF was proposed to address this issue by incorporating sparsity constraint into the basic NMF method and has demonstrated success in yielding succinct representations for face images[5,6]. In data clustering and classification problems, it is essential to consider the geometrical structure (manifold) of the data space; graph-regularized NMF (GNMF) was proposed[7] with the aim to find a new space representation such that the associated values at two data points are close to each other if they are connected in the graph. GNMF has demonstrated applications in pattern recognition [7] and tumor classification[8] . Other treatments of NMF, to name a few, include particular choices of the cost function (i.e., measure of the distance or divergence between the original and the factorizing

matrices) to reflect the underlying data generative model[9]; NMF in a Bayesian framework that treats the non-negativity constraint as priors and derive efficient approximation of the NMF factors as the posterior density based on a Gibbs sampler[10].

In the context of tumor subtype classification, the utility of GNMF was demonstrated on classification problems based on exome-level mutation data[8]. Although exome sequencing provides comprehensive characterization of coding mutations, it is likely that a large portion of mutations are passengers, as it was estimated that few mutations in a patient are drivers (e.g. ranging from 2 to 8)[11,12]. Such passengers, if included in clustering analysis, may obscure clinically and biologically important mutations. In a recent study, it revealed that using a panel of important genes can achieve superior classification than using the full set of (exome-level) mutations[13]. We hypothesize that a hybrid approach of sparse coding and GNMF will enable automatic selection of important genes that can aid tumor subtyping and interpretation of the underlying pathways.

To test our hypothesis, we propose a new method and evaluate it on multiple simulated mutation cohorts. The rest of the thesis is organized as follows. In Chapter 2, we give a brief overview on various NMF methods. In Chapter 3, we introduce a new formulation, called sparse network-regulated NMF, along with the update rule, convergence properties and stopping criteria. In Chapter 4, we evaluate the performance of the proposed sparse network-regulated NMF using simulation studies. In Chapter 5, we demonstrate real data applications using 13 major cancer types from The Cancer Genome Atlas (TCGA) dataset. In Chapter 6, we summarize the study and discuss future directions.

CHAPTER II

OVERVIEW OF NMF ALGORITHMS

2.1 Formulation

Given a data matrix X of dimensions $M \times N$ with nonnegative entries, NMF is the problem of finding a factorization

$$X \approx WH = \hat{X} \tag{1}$$

where W and H are matrices with non-negative entries and dimensions of $M \times K$ and $K \times N$, respectively. K is usually chosen such that $MK+KN << MN$, hence achieving dimension reduction of the data space X. The basis matrix W represents higher-level features of the data, and the coefficient matrix H represents loading coordinates of data points in the new space spanned by the basis vectors. The factorization (1) can be formulated as a minimization problem

$$\min_{W,H\geq 0} ||X - WH||_F^2 \tag{2}$$

where F indicates Frobenius norm, i.e., $||X||_F = \sqrt{\sum_{i,j} |x_{ij}|^2}$. More generally, the factorization (1) can be formulated as the following:

$$\min_{W,H\geq 0} d(X|WH) = D\left(X\middle|\hat{X}\right) = \sum_{m=1}^{M} \sum_{n=1}^{N} d(x_{mn}|\hat{x}_{mn}) \tag{3}$$

where $d(x|y)$ is a cost function. Popular choices are the squared Euclidean distance, Kullback-Leibler (KL) divergence, and Itakura-Saito (IS) divergence, which are defined respectively as[9]:

$$d_{EU}(x|y) = \frac{1}{2}(x-y)^2 \tag{4a}$$

$$d_{KL}(x|y) = x log \frac{x}{y} - x + y \tag{4b}$$

$$d_{IS}(x|y) = \frac{x}{y} - log \frac{x}{y} - 1 \tag{4c}$$

All cost functions are positive and have a single minimum at 0 when $x = y$.

2.2 Statistical Perspective

NMF can be recast from a statistical perspective. The choice of a certain cost function $d(\cdot \mid \cdot)$ to measure the fit between x_{mn} and $\hat{x}_{mn}$ implies certain assumptions about how x_{mn} is generated from $\hat{x}_{mn}$. It was already pointed out [9,14] that Euclidean, KL and IS NMF underlie the following generative models:

$$x_{mn} \sim N(\hat{x}_{mn}, \sigma^2) \qquad \text{EU-NMF} \qquad (5a)$$

$$x_{mn} \sim Pois(\hat{x}_{mn}) \qquad \text{KL-NMF} \qquad (5b)$$

$$x_{mn} \sim \Gamma(a, a/\hat{x}_{mn}) \qquad \text{IS-NMF} \qquad (5c)$$

where $N, Pois, \Gamma$ refer to the Gaussian, Poisson and Gamma distribution, respectively. In other words, NMF based on Euclidean distance underlies an additive Gaussian noise; NMF based on Kullback-Leibler divergence underlies a Poisson noise; NMF based on Itakura-Saito divergence underlies a multiplicative Gamma noise. Assuming these generative models, NMF algorithms can be seen as computing a maximum likelihood estimate (MLE) of the non-negative factorizing matrices. For example, NMF of X into WH based on Euclidean distance is equivalent to the MLE of the mean (WH) under a Gaussian model; when KL-divergence is used, NMF of X into WH is equivalent to the MLE of the mean under a Poisson model; when IS divergence is used, NMF of $X^T X$ (not X) into WH is equivalent to the MLE of W and H under a Gamma model. Of note, sometimes there may be interpretability ambiguity when the MLE of the Gaussian model does not guarantee non-negativity or when a Poisson model is used for real-valued data[9].

Bayesian treatment of NMF can easily avoid the interpretation ambiguity problem. In Bayesian frameworks, parameters are treated as priors and appropriate densities can be chosen in accordance with our beliefs about the parameters. Moreover, Bayesian NMF can utilize powerful Bayesian tools for estimating the posterior density and obtain uncertainty estimate. In the work[15]

of Schmidt et al. (2009), the proposed Bayesian NMF features 1) an exponential distribution as the prior to ensure non-negativity; 2) an efficient Gibbs sampler to approximate the posterior density of the NMF factors; and 3) order selection via marginal likelihood estimation[15].

2.3 Algorithms

Non-negative matrix factorization is a nonlinear optimization problem. The objective function (2) is convex in W or H but not both. Hence it is unrealistic to expect an algorithm to find the global optimum. Various numerical approaches have been proposed, including multiplicative algorithms[1,4,7], alternating least squares algorithms (ALS)[6,16], to name a few. Given the large number of parameters to estimate (KN+KM), convergence to a local minimum is only guaranteed for some algorithms. Multiplicative update algorithms are well supported by convergence theories, and have been popular and standard methods.

In the following, we briefly introduce basics on the multiplicative algorithms. In our proposed method, we also use a multiplicative update rule.

2.3.1 Multiplicative Algorithms

Lee and Seung first proposed to use multiplicative iterative algorithm to search for the local minima of NMF[1]. The multiplicative approach has been well described[15]: it updates each parameter by multiplying its value at previous iteration by the ratio of the negative and positive parts of the derivative of the criterion w.r.t. this parameter, namely, $\theta \leftarrow \theta . \frac{[\nabla f(\theta)]_-}{[\nabla f(\theta)]_+}$ where $\nabla f(\theta) = [\nabla f(\theta)]_+ - [\nabla f(\theta)]_-$ and the summands are both nonnegative. This ensures non-negativity of the parameter updates, given that the initialization values are non-negative. A fixed

point of the algorithm implies either $\nabla f(\theta^*) = 0$ or $\theta^* = 0$. Under EU-NMF it leads to the following updates[1]

$$w_{ij} \leftarrow w_{ij} \frac{(XH^T)_{ij}}{(WHH^T)_{ij}} \quad (6a)$$

$$h_{ij} \leftarrow h_{ij} . \frac{(W^T X)_{ij}}{(W^T WH)_{ij}} \quad (6b)$$

where w_{ij} and h_{ij} are the entries of matrices W and H. It has been pointed out that the update equations of EU-NMF, KL-NMF and IS-NMF can be re-written into unified coupled equations with one parameter to indicate the specific NMF variant[9]

Of note, given the multiplicative iteration, the W and H matrices may not be sparse, containing lots of low values close to zero but not exactly 0. On the other hand, once an element becomes zero, it will remain at 0. Locking at '0's too early may lead to a poor solution path.

2.4 Regularization

In regression, various regularization methods have demonstrated their utility for variable selection, especially in the 'high dimensional data small sample size' scenarios. Lasso penalizes a least squares regression by the sum of the absolute values (L_1-norm) of the coefficients[17]. The form of this penalty encourages sparse solutions (i.e., many coefficients equal to 0). Elastic net combines both lasso and ridge penalties and has demonstrated a superior performance than lasso while enjoying a similar sparseness of representation[18]. Elastic net encourages a grouping effect, where strongly correlated predictors tend to be in (or out) the model together and the regression coefficients tend to be equal[18]. Fused lasso penalizes the L_1-norm of both the coefficients and their successive differences, leading to sparse and smooth coefficients for features ordered in some particular way[19]. Fused lasso has demonstrated its utility in analysis of genomic copy

number data. Both elastic net and fused lasso could be thought of as special cases of group lasso, which provides a more flexible framework to incorporate constraints on pre-defined groups of variables[20].

Graph-regularized NMF is similar to group lasso in spirit. This is done by constructing a nearest neighbor graph, incorporating the graph structure into the objective function of matrix factorization[7]:

$$min_{W\geq 0, H\geq 0} \left|\left|X - WH\right|\right|_F^2 + \lambda\, trace(W^T LW) \quad (7)$$

where $\|\cdot\|_F$ denotes the matrix Frobenius norm, λ is the regularization parameter, L is the graph Laplacian[21] of a k-nearest neighbor graph. In the context of tumor subtyping, each basis vector of W represents a 'metagene', a collection of genes as the functional unit underlying a subtype. This penalization encourages similar weights for genes if these genes are known to connect or interact in a common pathway[8].

CHAPTER III

SPARSE NETWORK-REGULATED NMF

With the same motivation as in penalized regression, we propose a new formulation of NMF to allow the control of both sparsity and graph structure on the basis matrix. We call this method sparse network-regulated NMF (sgNMF), implemented in two steps.

Step One

Step one learns the basis representation under the graph constraint. We used the iterative algorithm proposed by Cai *et al.*[7] to find solutions W and H. Briefly, we re-write the equation (7), apply Lagrange multiplier, take partial derivative w.r.t. W an H, and apply Karush-Kuhn-Tucker (KKT) condition to obtain the following update steps:

$$w_{ij} \leftarrow u_{ij} \frac{(XH^T + \lambda AW)_{ij}}{(UHH^T + \lambda DW)_{ij}} \quad (8a)$$

$$h_{ij} \leftarrow h_{ij} \frac{(X^T W)_{ij}}{(H^T W^T W)_{ij}} \quad (8b)$$

This is a multiplicative update rule. The proof of convergence has been demonstrated[7]. The stopping criterion was set when the absolute difference in values of the objective function between two consecutive iterations is less than 0.1.

Step Two

Step two applies a lasso-like constraint to modify W and solves for H that minimizes the reconstruction error. Specifically, given the solution W obtained from step one, all entries below threshold α ($\alpha > 0$) are set to exactly 0, and the rest entries subtract α. Denote the resulting

matrix as W_1. The coefficient matrix H_1 is solved by least squares and then set the negative values to zero to ensure non-negativity:

$$\begin{cases} H = (W_1^T W_1)^{-1} W_1^T X \\ H_1 \Rightarrow H \geq 0 \end{cases} \quad (9)$$

It is our interest to test whether the clustering based on the new H_1 will improve compared to the clustering based on the original H.

We are aware that a potential drawback of the proposed sgNMF is that it does not simultaneously control sparsity and graph structure. However, a lasso-like constraint is infeasible to realize given a multiplicative update rule that scales the results, giving lots of low values close to zero rather than setting them to exactly 0. Then why not use other update algorithms? The reason we chose the multiplicative algorithm over other algorithms is that, other algorithms such as gradient descent algorithms and alternating least squares (ALS) algorithms suffer from their own problems when used to obtain numerical solutions of NMF. The difficulty of gradient descent algorithms comes in choosing the step size to ensure convergence and non-negativity; the ALS algorithms suffer from convergence partially due to the ad hoc implementation of nonnegativity[22]. Furthermore, penalization is hard to implement within ALS. For these reasons, we chose a compromised strategy to use multiplicative rule to control graph structure followed by a sparseness correction to control sparsity.

CHAPTER IV

SIMULATION STUDIES

4.1 Simulation Setup

We choose to use simulation to assess the ability of sgNMF to recover true subtypes from somatic mutation profiles. For most cancers, the true subtypes are unknown. For a few cancer types that do have known subtypes such as colon rectal cancer and breast cancer, the subtype classifications are based on gene expressions[23] [http://sagebase.org/case-studies/colorectal-cancer-subtyping-consortium/]. The substantial difference in the two molecular data types makes it inappropriate to set one as the golden standard for the other to compare. Therefore, we set out to use synthesized data to carry out systematic evaluations of the sgNMF method. We simulate somatic mutation cohorts as follows.

4.1.1 Mutation Cohorts

For each sample, we indicated a gene as 1 if it mutated and zero otherwise. The number of patients per cohort and the number of mutations per patient follow the ovarian cancer mutation dataset from TCGA. We first permuted for each patient the mutation profile while keeping the per-patient mutation frequency invariant. This was to generate a background with no subtype signal. Then, a network-based signal was added to the patient-by-mutation matrix as follows. First, we established a set of network modules (i.e., connected components enriched for edges shared within modules) in an input network using a fast spectral clustering algorithm. The input network we chose is HumanNet[24] v.1 with top 10 percent of the most confident edges and 11 nearest-neighbors; we used the R package 'mclust' for module detection based on the fast spectral clustering. Of the 120 modules detected, we focused on modules of size below 150

genes. Next, we divided the patient cohort randomly into four equal-sized subtypes (four was selected as reasonable owing to the four expression-based subtypes that have been identified for glioblastoma, ovarian and breast cancers[11,25-27]). Each subtype was assigned a network module that had size *s* ranging from 10 to 150 genes; these modules are denoted as driver modules. In order to estimate the appropriate size of driver modules (i.e., cancer pathways), we examined the known cancer pathways in the NCI-Nature pathway interaction database[28]. The database of NCI-nature has curated 136 cancer pathways with sizes ranging 2-139 and a medium of 34.

We assume zero overlap (no shared genes) between these driver modules. For each patient, we reassigned a fraction of the patient's mutations f to genes covered by the driver module that characterizes the patient's subtype. A plausible range for the number of driver mutation in a tumor was recently proposed to be between 2 to 8 driver mutations[12]. We note that a fraction of 4% corresponds to between 2 and 9 mutations (median of 3), which is consistent with the aforementioned estimate.

4.1.2 Controlling Subtype Signals

We varied the strength of subtype signals by controlling the 'cancer pathways' that determine the subtypes. This was done by controlling the size and connectivity of the modules underlying the subtypes (Table 1). Module connectivity is measured by graph density, i.e., the number of edges divided by the number of edges if the graph is fully connected. A highly connected module is expected to have stronger signals than a module that is less connected. We divided the modules based on the quantile of the connectivity level; within each level, module sizes were chosen to generate either an equal size distribution or a step-down size distribution. Table 1 lists multiple

simulated cohorts whose subtypes were assigned based on modules of different connectivity levels and sizes.

4.2 Network-based Stratification

We performed subtype classification of the pseudo tumors using the network-based classification (NBS) framework developed by Hofree *et al.*[8]. Tumor classification was done by applying the original NBS (with GNMF) as well as the modified NBS (with sgNMF). The resulting partitions were compared with the true simulated subtypes to compare the two approaches and assess the effect of the sparse constraint. We used adjusted random index[29] to compare two partitions.

NBS was recently proposed to overcome the challenge in classification on somatic mutations by leveraging information provided in protein-protein interaction networks (PPI)[8]. Briefly, NBS uses label propagation on PPI to assign higher values to non-mutated genes that are closer to genes (in PPI) that harbor mutations. This guilt-by-association principle governed by genetic networks has many applications for biological discovery utilizing prior knowledge. For somatic mutations in genes, this principle fits well with the underlying biology: driver genes are often interacting directly or indirectly in common pathways and mutations in different genes in the same pathway are likely to cause genetically similar tumor[30]. NBS has been applied on several cancers using exome-level mutation data and showed improved association of subtypes with clinical outcomes than using mRNA data. In general NBS provides a unified framework to further investigate tumor subtyping by integrating somatic mutations with biological networks[8,13].

4.2.1 Network Propagation

Let X_0 be the initial gene × patient matrix ($M \times N$), and A be the symmetric adjacency matrix representing gene-to-gene interaction network ($M \times M$). The network propagation process is carried out by the following iterative algorithm[8]:

$$X_{t+1} = a \cdot A \cdot X_t + (1 - a) \cdot X_0 \qquad (10)$$

We set $a = 0.7$ as previously described [8]. The propagation function was run iteratively until F_t converges ($|F_{t+1} - F_t| < 0.001$). Following the propagation, quantile normalization was applied to F_t to ensure each patient follows the same distribution for the smoothed mutation profile. We use F to denote the final normalized and smoothed mutation matrix.

4.2.2 Network-regularized NMF

This step can be deemed as an application of GNMF on mutation data (see equation 7). The value K controls the dimension reduction, and we used K=3,4,5,6 in this study. L is the graph Laplacian of a k-nearest-neighbor network. We chose k=11 as previously described[8]. λ is the regularization parameter and the value was set by parameter tuning (see below). The iteration was run until the objective function converges ($|X_{t+1} - X_t| < 0.1$).

4.3 Tuning Parameters

Selection of the regularization parameters should be very careful to gain a balance between sparseness and discrimination. During iterations, under the multiplicative rule, once an element in W or H becomes 0, it must remain 0. Thus, a strong penalty λ in equation (2) may cause a solution path to settle early to a poor solution. The value of α controls the degree of sparseness in the basis matrix. The larger the α, the more succinct the W representation is and the larger the

departure from the original solution. This may cause large residuals in approximating the original matrix. We first tried a coarse combination of λ and α covering a wide range. The initial results suggest the use of a more focused region with a finer spacing. Specifically, we used a grid of $\lambda \times \alpha$ with a total of 105 combinations, whereas λ =10, 5, 4, 3, 2,1, 0.8, 0.7, 0.6, 0.5, 0.2, 0.1, 0.05, 0.01, 0.001 and α =0.1, 0.08, 0.005, 0.02, 0.01, 0.005, 0.001 (most elements in W is small). We used classification accuracy in simulated data to guide the choice on parameter values.

4.4 Results

4.4.1 Proof of Concept

As a proof of concept, we first tested our method on a cohort with strong signals determining the subtypes. Specifically, we picked four modules to represent the cancer pathways with size <50 and high connectivity (top 1, 2, 3, 12, respectively) among the 120 modules. Table 2 lists the size and connectivity of the modules characterizing the subtypes. In these scenarios, the classification accuracy was mostly >98% if λ is small (close to 2). The sparseness constraint also produced sparse basis vectors. With α=0.05, the metagene corresponding to each subtype contains 40, 36, 41, 24 non-zero values, respectively. These non-zero values characterize the major contributing genes for each metagene. There is a good overlap between the detected metagenes and the cancer pathways underlying the subtypes (Table 2). For subtype II-IV, each 'cancer pathway' was totally covered by the corresponding metagene. For subtype I, the metagene failed to cover the whole 'cancer pathway' but was also enriched with the 'cancer genes' (37/40). In short, the metagenes demonstrated good sensitivity and specificity to recover the underlying cancer pathways. In contrast, without the sparse constraint, the metagenes are dense, each containing

more than 14,000 non-zero elements (out of 16179), although the top genes are still in the cancer pathways.

4.4.2 Tuning Parameters

We consistently observed that, the optimal or sub-optimal classification accuracy occurred in the region [0.001,1]×[0.001, 0.05] of the $\lambda \times \alpha$ parameter space. Furthermore, the change in the classification accuracy over this parameter subspace is relatively small (min, median and max of the accuracies were shown in Table 3). For most cases, the *max* - *min* difference is less than 0.1, regardless of whether sparsity was imposed or not (Table 3). In terms of the relative contribution, we observed a more decisive effect of λ on the classification accuracy while only a minor effect of α. This is not surprising, as our two-step design restricts the role of α in step one, and the need to balance sparsity and good approximation favors small α in step two, leading to small changes in the resulting coefficient matrix. Based on the simulated data we chose setting λ=1 and α=0.05 in other analyses.

4.4.3 Pathway Connectivity and Classification Accuracy

We observed a positive relationship between classification accuracy and connectivity of the pathway sets. This is expected, because a densely connected pathway enables the effect of mutations to propagate through the common pathway via gene-gene interaction. Thus, the common pathway will be picked and the (pseudo) tumors are clustered together despite that the driver mutations are from different genes. As shown in Table 3, once the connectivity is above 0.1 for all modules (cohort 16), the signal will be strong, resulting in high classification accuracy (>98%); when the connectivity level is close to 0.05, often times the cohort achieved sufficiently

good classification accuracy (>86%). This occurred regardless of the sparsity constraint (Table 3). We also commented that graph connectivity is not the single factor determining the subtype signals. For example, cohort 6 and 7 had similar connectivity but differed substantially in classification accuracy (Table 3). This is probably because cohort 6 has larger modules. By definition of graph density, it is much harder for a large subnetwork than a small one to reach certain edge density, and for the same connectivity larger modules usually have more effective label propagation. Another example that is more difficult to explain is by comparing cohort 11 and 12. Now cohort 11, which had smaller modules of lower connectivity, achieved much higher classification accuracy than cohort 12 (Table 3).

4.4.4 Effect of Sparseness Constraint

We observed a limited effect of sparseness constraint on the classification accuracy across all simulations. Imposing sparsity introduced a small amount of disturbance to the original solution W, and more often led to a classification that is slightly better (Table 4). It was less often that the final classification became worse or remained unchanged, especially when the subtype signals are strong and the original classification is already good (see cohort 13, 14, 16 in Table 4).

In summary, our simulation studies suggest the two following points. First, the GNMF method works as expected, the sparse constraint has mixed impacts on the classification accuracy and often times led to a slightly better solution. Since NMF algorithms produce local optimums by design due to the non-convex property, the run-to-run variations should exceed the variations by imposing the sparseness constraint. In practical use, we recommend to use GNMF combined with consensus clustering (see below) to achieve robust classification and applying the

thresholding to obtain sparse representation of the basis vectors to aid interpretation of pathways. This can be thought of as a modified NBS strategy.

CHAPTER V

APPLICATIONS TO REAL DATA

We applied the modified analysis strategy as recommend at the end of last chapter to 13 tumor types from TCGA cohorts. In addition to clustering patterns, we also perform survival analysis to identify clinically relevant subtypes. Finally, we used CRC as an example to apply sparse coding to obtain a list of genes most responsible for the detected subtypes.

5.1 Mutation Data

We collected mutation data on solid tumors of 13 major cancer types profiled by exome-sequencing and investigated previously[30] (http://cancergenome.broadinstitute.org). We focus on mutations that are non-synonymous, occurred on splice sites or stop codons (termed "functional" mutations). In contrast, the non-functional mutations refer to synonymous mutations and mutations in intronic or intergenic regions. Samples with fewer than 6 functional mutations in exomes were discarded. Finally, we are left with a total of ~4000 samples (Table 4).

5.2 Consensus Clustering

In order to achieve robust clustering, we used consensus clustering[31] to generate the final clustering of patients. Specifically, we ran network-regulated NMF using a random sample without replacement of 80% patients to construct a clustering, and repeat this process 50 times. The collection of 50 clustering results was used to construct the similarity matrix, which records the frequency with which each pair of patients was observed to share the same membership among all replicates. Hierarchical clustering with average linkage was generated based on the similarity matrix using the R "NMF" package.

5.3 Survival Analysis

Survival analysis was performed using the R "survival" package. Kaplan-Meier survival curves were plotted for each NBS subtype and log-rank tests were performed to test the association of subtypes with survival.

5.4 Results

Figure1-3 show the clustering pattern of 8 cancer types (BRCA, CRC, GBM, HNSC, KIRC, LUAD, OV, UCEC) for different ranks (K=2,3,4,5). The clustering patterns are clear, and all tumor types have at least one clear clustering pattern for certain K. Other five cancers types had less clear clustering patterns and were not shown.

Figure 4 shows the NBS subtypes and survival for six cancer types (BRCA, CRC, GBM, HNSC, KIRC, UCEC). The p-values from log-rank tests are shown for each cancer type. Despite the clear pattern of clustering, none of the p-values are significant. The best case is GBM, whose p-value is marginal (0.03). After correction for multiple testing, none of the cancer types had subtypes that are associated with survival. The reasons are several folds. First, subtypes based on molecular profiles do not necessarily result in distinct survival or other clinical demonstration. Different pathways could present similar clinical symptoms when they are disturbed. Second, somatic mutations only capture one source of genomic aberrations. Each of the multiple data types such as gene expression, DNA methylation and copy number etc., each contains its own unique information, and solely relying on mutation data may miss other important pieces of information. Third, the survival data used in the analysis may not be of good quality. For example, the follow-up times were generally short in the TCGA CRC cohort, which affects the

quality of the survival data. Finally, the samples were collected and profiled at one time point and the patients were followed up for a period of time during which, some medicine or therapies were administered such that the survival will not reflect the molecular profiles when they were collected.

CHAPTER VI

SUMMARY

In summary, we proposed a new formulation to incorporate sparse coding into the graph-regulated NMF in order to improve accuracy of classifying tumor subtypes. Due to the lack of well-established subtypes for most cancer types, we set out to simulate mutation cohorts with underlying driver cancer pathways. We evaluated new method over a spectrum of simulated mutation cohorts varying in signals of determining the subtypes. We found that the sparseness constraint more often led to slightly better solutions. Such limited effect is in part due to the two-step design of our algorithm in forcing the sparse representation. Furthermore, in tuning the regularization parameters, we identified a parameter region in which the classification accuracy was qualitatively better for all simulated cohorts we examined. We recommend using the original NBS for subtype detection and sparse coding for interpretation of the pathways underlying the detected subtypes. For illustration purpose, we applied this analysis strategy to several tumor types from TCGA, and identified clustering patterns for eight cancer types including BRCA and CRC. Our association analysis revealed that the majority of the NBS subtypes are not associated with survival. We provided potential reasons why no association between the NBS subtypes and survival were detected. Finally, using CRC as an example, we provide a list of genes most responsible for the subtypes detected using sparse coding.

REFERENCES

1. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature.* Oct 21 1999;401(6755):788-791.

2. Gaussier E, Goutte C. Relation between PLSA and NMF and implication. *Research and Development in Information Retrieval, Proceedings of the International SIGIR Conference.* 2005:601-602.

3. Schmidt MN, Olsson RK. Single-channel speech separation using sparse non-negative matrix factorization. *Spoken Language Processing, ISCA International Conference.* 2006.

4. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences of the United States of America.* Mar 23 2004;101(12):4164-4169.

5. Li SZ, Hou X, Zhang H, Cheng Q. Learning spatially localized, parts-based representation. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* 2001;1:207-212.

6. Hoyer PO. Non-negative Matrix Factorization with Sparseness Constraints. *Journal of Machine Learning Research* 2004;5:1457–1469.

7. Cai D, He X, Wu X, Han J. Non-negative matrix factorization on manifold. *8th IEEE International Conference on Data Mining.* 2008:63-72.

8. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. *Nature methods.* Nov 2013;10(11):1108-1115.

9. Fevotte C, Cemgil AT. Nonnegative matrix factorizations as probabilistic inference in composite models. *17th European Signal Processing Conference.* 2009(1913-1917).

10. Schmidt MN, Winther O, Hansen LK. Bayesian non-negative matrix factorization. *Independent Component Analysis and Signal Separation Lecture Notes in Computer Science.* 2009;5441:540-547.

11. Kandoth C, McLellan MD, Vandin F, et al. Mutational landscape and significance across 12 major cancer types. *Nature.* Oct 17 2013;502(7471):333-339.

12. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW. Cancer genome landscapes. *Science.* Mar 29 2013;339(6127):1546-1558.

13. Zhong X, Yang H, Zhao S, Shyr Y, Li B. Network-based stratification analysis of 13 major cancer types using mutations in panels of cancer genes. BMC *genomics.* Jun 11 2015;16 Suppl 7:S7.

14. Fevotte C, Bertin N, Durrieu JL. Nonnegative Matrix Factorization with the Itakura-Saito Divergence:With Application toMusic Analysis. *Neural Computation.* 2009;21:793–830.

15. Dikmen O, Fevotte C. Maximum Marginal Likelihood Estimation for Nonnegative Dictionary Learning in the Gamma-Poisson Model. *IEEE Transactions on signal processing.* 2012;60(10).

16. Paatero P, Tapper U. Positive matrix factorization: A nonnegative factor model with optimal utilization of error-estimates of data values. *Environmetrics* 1994;5(2):111–126.

17. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B.* 1996;58(1):267-288.

18. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B.* 2005;67(2):301–320.

19. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B.* 2005;67(1):91-108.

20. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B.* 2006;68(1):49-67.

21. Luxburg U. A Tutorial on Spectral Clustering. *Statistics and Computing.* 2007;17(4):395-416.

22. Berry MW, Browne M. Algorithms and Applications for Approximate Nonnegative Matrix Factorization. *Computational Statistics and Data Analysis, .* 2007;52(1):155–173.

23. Parker JS, Mullins M, Cheang MC, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology.* Mar 10 2009;27(8):1160-1167.

24. Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome research.* Jul 2011;21(7):1109-1121.

25. Frampton GM, Fichtenholtz A, Otto GA, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nature biotechnology.* Nov 2013;31(11):1023-1031.

26. K.A. H, Yau C, Wolf DM, Cherniack AD, Tamborero D. Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin. *Cell.* 2014;158(4):929-944.

27. Kreeger PK, Lauffenburger DA. Cancer systems biology: a network modeling perspective. *Carcinogenesis.* Jan 2010;31(1):2-8.

28. Schaefer CF, Anthony K, Krupa S, et al. PID: the Pathway Interaction Database. *Nucleic acids research.* Jan 2009;37(Database issue):D674-679.

29. L. H, P. A. Comparing Partitions. *Journal of the Classification.* 1985;2:193-218.

30. Lawrence MS, Stojanov P, Mermel CH, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature.* Jan 23 2014;505(7484):495-501.

31. Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning.* 2003;52:91–118.

TABLES

Table 1: Size and connectivity of the simulated cancer pathways characterizing the subtypes in the simulated mutation cohorts.

Simulated Cohorts[a]	Cancer Pathway							
	Size				Connectivity			
	I	II	III	IV	I	II	III	IV
1	127	131	153	87	0.003	0.002	0.002	0.001
2	59	68	87	127	0.003	0.002	0.001	0.003
3	55	52	53	53	0.01	0.009	0.007	0.004
4	77	62	73	73	0.01	0.005	0.004	0.003
5	37	33	34	32	0.014	0.009	0.007	0.006
6	104	116	106	117	0.05	0.049	0.047	0.045
7	70	80	89	81	0.05	0.046	0.033	0.021
8	50	56	62	69	0.047	0.04	0.036	0.032
9	32	38	31	36	0.044	0.034	0.019	0.016
10	151	94	62	26	0.05	0.042	0.036	0.015
11	104	80	56	36	0.05	0.046	0.04	0.016
12	108	92	89	99	0.054	0.052	0.053	0.068
13	60	65	67	68	0.101	0.063	0.058	0.054
14	40	42	40	48	0.099	0.098	0.097	0.09
15	108	46	35	60	0.054	0.118	0.19	0.101
16	49	31	35	14	0.12	0.17	0.19	0.43

[a] Cohorts were divided into 4 blocks by connectivity:1-2, connectivity$\leq$0.003; 3-5, 0.003<connectivity $\leq$0.01; 6-11, 0.01<connectivity$\leq$0.05; 12-16, connectivity>0.05.

Table 2: Simulated 'cancer pathways' and the pathways ('metagenes') detected by sparse coding combined with the NBS approach

Subtype	Cancer Pathway		Metagene	Overlap between metagene and cancer pathway
	Size	Connectivity	Size	
I	49	0.12	40	37
II	31	0.17	36	31
III	35	0.19	41	35
IV	14	0.43	24	14

Classification accuracy of 100% using λ =1 and α =0.05

Table 3: A region of regularization parameters identified to have relatively stable classification accuracy

Simulated Cohorts	Classification Accuracy	
	w/o sparsity constraint	w. sparsity constraint
	min, median, max[a]	min, median, max[a]
1	0.044 0.058 0.066	0.028 0.074 0.079
2	0.114 0.12 0.123	0.082 0.112 0.126
3	0.413 0.557 0.631	0.299 0.542 0.685
4	0.172 0.268 0.301	0.149 0.288 0.338
5	0.88 0.914 0.923	0.206 0.923 0.94
6	0.906 0.948 0.948	0.804 0.991 0.991
7	0.501 0.559 0.57	0.418 0.574 0.589
8	0.532 0.861 0.948	0.245 0.904 0.948
9	0.889 0.914 0.939	0.426 0.939 0.948
10	0.52 0.568 0.578	0.371 0.547 0.561
11	0.858 0.898 0.914	0.719 0.922 0.931
12	0.61 0.636 0.914	0.603 0.645 0.879
13	1 1 1	0.956 1 1
14	0.991 1 1	0.965 1 1
15	0.574 0.601 0.621	0.56 0.586 0.6
16	0.982 1 1	0.991 1 1

[a] The min, median and max of the classification accuracy resulting from 50 parameter combinations: $\lambda\epsilon\{1, 0.8, 0.7, 0.6, 0.5, 0.2, 0.1, 0.01, 0.001\}$ and $\alpha\epsilon$ $\{0.05, 0.02, 0.01, 0.005, 0.001\}$. w/o, without; w., with.

Table 4. Classification efficiency with and without sparseness constraint for various parameter combinations[a] in the simulation data

Simulated Cohorts	Difference in Classification Accuracy (Sparse - Original)					Frequency[a]		
	Min	1st Qu.	Median	3rd Qu.	Max	Sparse is Better	Original is Better	Equal
1	-0.027	0	0.014	0.023	0.041	50%	50%	0%
2	-0.034	-0.018	-0.008	0.002	0.045	29%	71%	0%
3	-0.334	0.005	0.044	0.103	0.165	56%	44%	0%
4	-0.1	-0.013	0.015	0.057	0.111	54%	46%	0%
5	-0.709	-0.007	0.013	0.026	0.068	54%	40%	6%
6	-0.144	0	0.035	0.043	0.094	45%	46%	10%
7	-0.137	-0.003	0.01	0.023	0.062	70%	28%	2%
8	-0.447	0	0.012	0.059	0.184	55%	42%	3%
9	-0.478	-0.006	0.004	0.032	0.059	44%	45%	11%
10	-0.165	-0.093	-0.021	-0.007	0.021	27%	73%	0%
11	-0.203	-0.027	0.009	0.041	0.082	46%	52%	2%
12	-0.082	-0.01	0.009	0.019	0.059	53%	46%	1%
13	-0.082	0	0	0	0	24%	31%	45%
14	-0.044	0	0	0.009	0.009	40%	30%	30%
15	-0.035	-0.036	0	0.011	0.036	49%	48%	4%
16	-0.062	-0.009	0	0	0.018	30%	31%	39%

[a] Parameter combinations: $\lambda \epsilon \{1, 0.8, 0.7, 0.6, 0.5, 0.2, 0.1, 0.01, 0.001\}$ and $\alpha \epsilon \{0.05, 0.02, 0.01, 0.005, 0.001\}$, a total of 50 combinations.

Table 5: Mutation data and clinical data from TCGA cohorts

Cancer[a]	Sample size (# mutations≥6)	Survival	Grade	Stage
BLCA	99	X	X	
BRCA	849	X		
CRC	233	X		
ESO	140	X	X	Missing 67%
GBM	288	X		
HNSC	372	X	X	X
KIRC	414	X	X	
LUAD	391	X		
LUSC	176	X		
MEL	118			
MM	200			
OV	313	X	X	X
UCEC	247	X	X	X
Total	4008			

[a]BLCA-Bladder urothelial carcinoma; BRCA-Breast invasive carcinoma; CRC-Colorectal carcinoma; ESO-Esophageal adenocarcinoma; GBM-Glioblastoma multiforme; HNSC-Head and neck squamous cell carcinoma; KIRC-Kidney renal clear cell carcinoma; LUAD-Lung adenocarcinoma; LUSC-Lung squamous cell carcinoma; MEL-Melanoma; MM-Multiple myeloma; OV-Ovarian serous cystadenocarcinoma; UCEC-Uterine corpus endometrial carcinoma

X marks availability of the clinical data type (survival, tumor stage, tumor grade).

Table 6: Top genes responsible for discriminating the CRC NBS subtypes

NBS subtype[a]	Gene
I	KRAS, APC, RASSF2, RASGRP2, RAP1GDS1, PIK3CG, PTHLH, SHOC2, PIK3K5, WNT1
II	TTN, MYOM2, MYBPC3, SYNE1, NEB, ANKRD23, ANKRD1, MYPN, SCN10A, SUNC1 ANK1, CACNA1A
III	TP53, PARC, E4F1, RCN2, SSTR3, ANKRD2, CABLES1, TP53BP2, EI24, APC

[a] assume K=3

FIGURE LEGEND

Figure 1: The clustering pattern of BRCA and CRC for rank K=2,3,4,5.

Figure 2: The clustering pattern of GBM, HNSC and KIRC for rank K=2,3,4,5.

Figure 3: The clustering pattern of LUAD, OV and UCEC for rank K=2,3,4,5.

Figure 4: Subtypes and the associated survival for BRCA, CRC, GBM, HNSC, KIRC and UCEC. P-values from the tests of association between subtype and survival were shown.

FIGURES

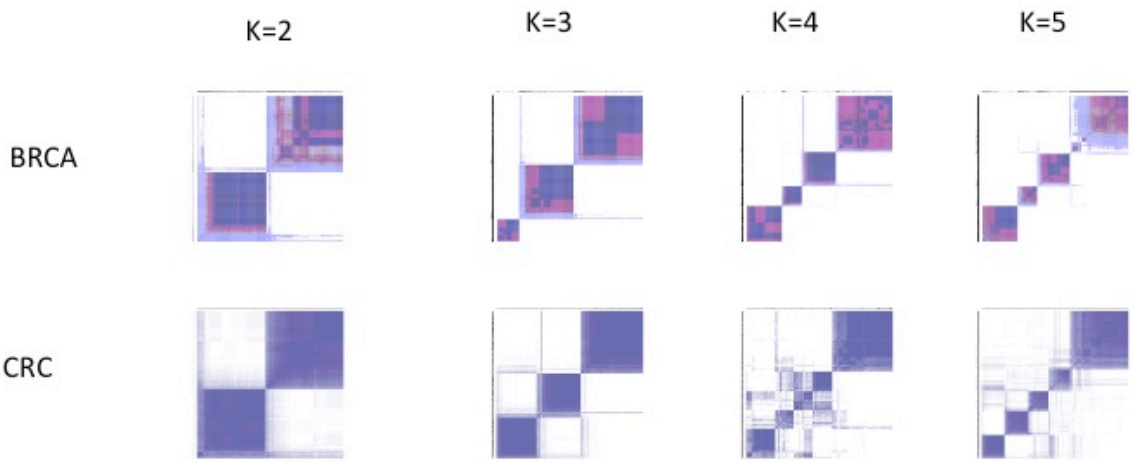

Figure 1

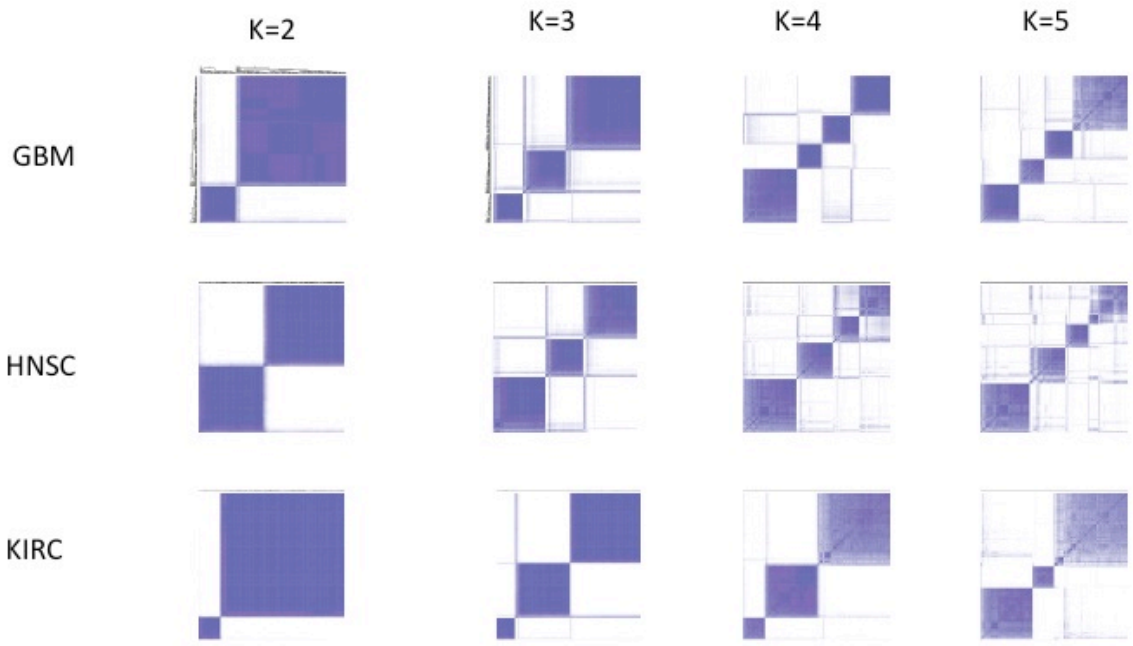

Figure 2

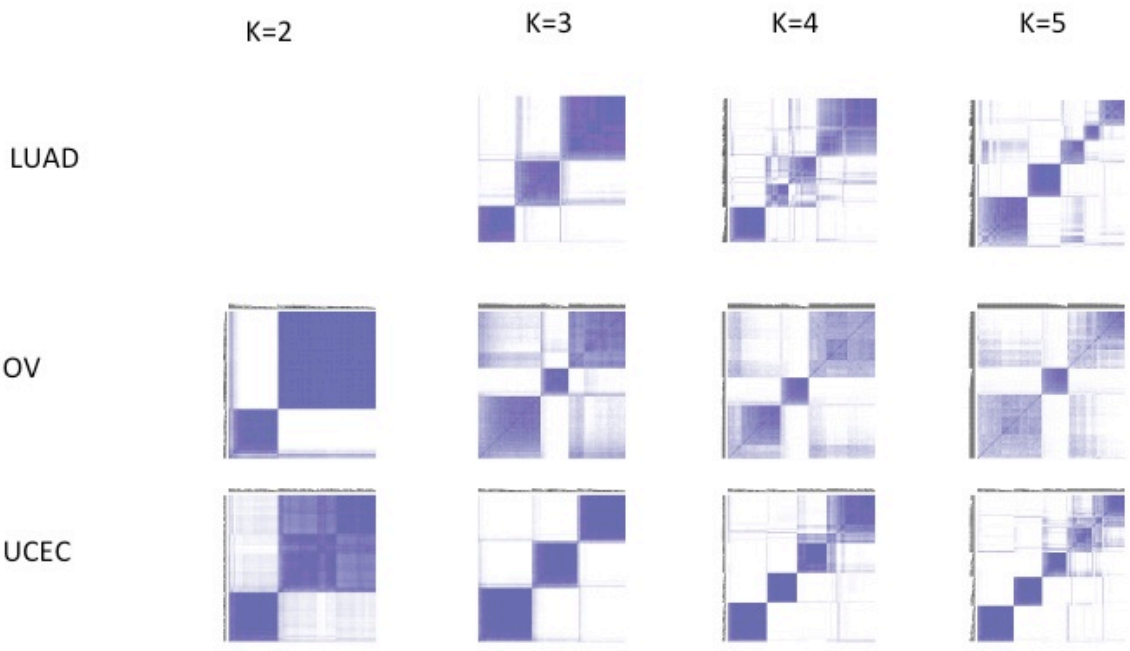

Figure 3

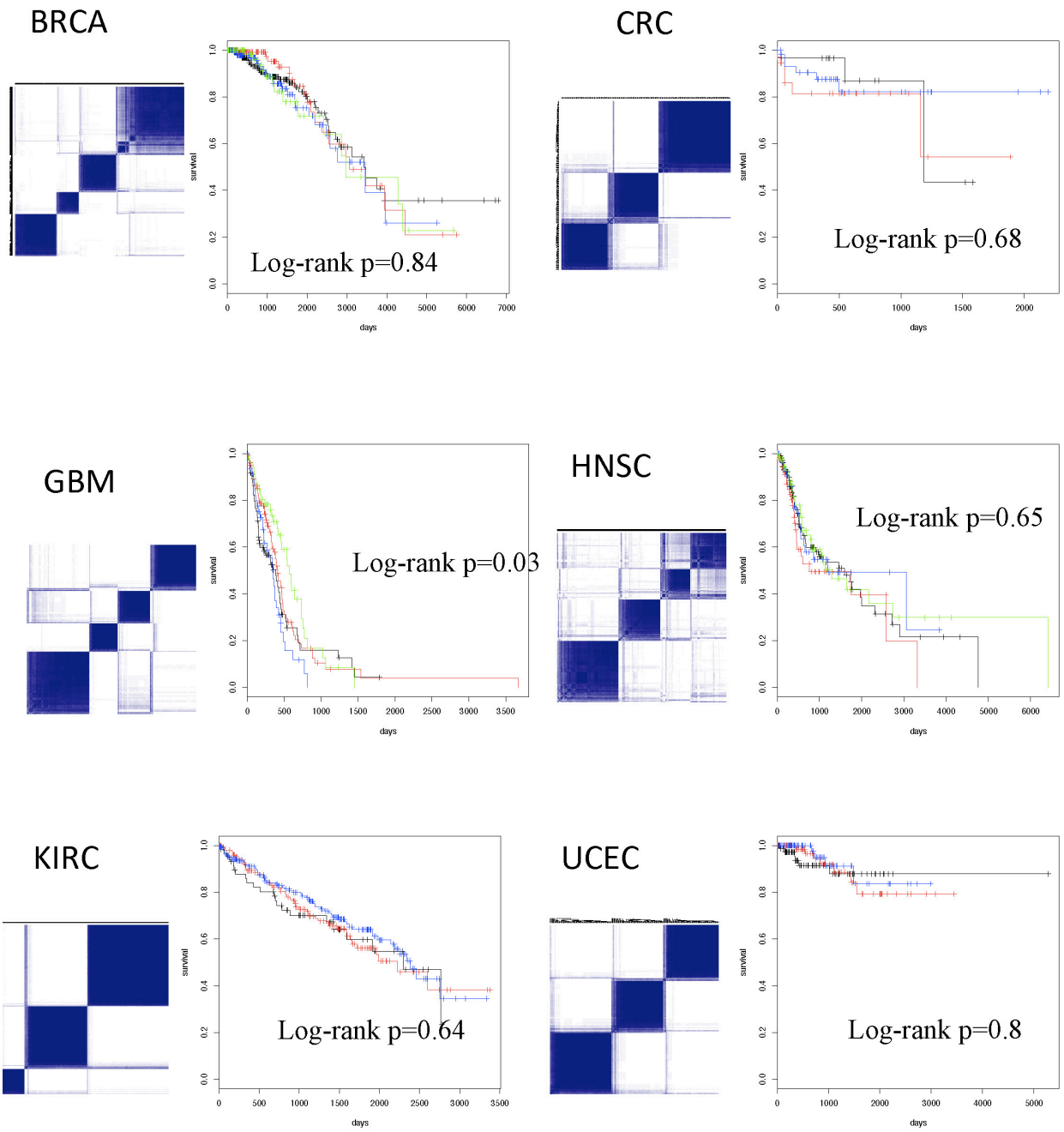
BRCA
Log-rank p=0.84
CRC
Log-rank p=0.68
GBM
Log-rank p=0.03
HNSC
Log-rank p=0.65
KIRC
Log-rank p=0.64
UCEC
Log-rank p=0.8
survival
days

Figure 4.