EXPECTATION-MAXIMIZATION BASED MIXTURE-MODEL EXPLOITING PATHWAY KNOWLEDGE FOR CANCER HETEROGENEITY

A Thesis

by

RAJAN KAPOOR

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Aniruddha Datta
Committee Members,	Ulisses Braga Neto
	Sunil Khatri
	Rabi Mahapatra
Head of Department,	Miroslav Begovic

May 2017

Major Subject: Electrical Engineering

ABSTRACT

Cancer cells are much more prone to mutations than normal cells, generating over time, more genetic variants of themselves within the tissue. Drugs designed for one variant might not work as intended for other variants. As such, effective drug design requires estimation of proportion of various cancer subpopulations.

In this work, a mixture model based approach with expectation maximization is proposed for determination of cancer heterogeneity. We exploit the pathway knowledge collected by biologists over time to surpass the limitations of identifiability shown by mixture models. Also in cases where Expectation-Maximization converges to more than one solution, pathway knowledge is used to break the tie by defining an error metric. Finally, using experimental data, changes in composition of the mixture over time are estimated using the model. The approach can also be used to compare the effectiveness of different drugs on a heterogeneous cancer tissue by observing the response over time.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a thesis committee consisting of Professors Aniruddha Datta, Ulisses Braga Neto and Sunil Khatri of the Department of Electrical and Computer Engineering and Professor Rabi Mahapatra of the Department of Computer Science and Engineering.

The data analyzed for Chapter 4 was provided by Drs. Chao Sima, Jianping Hua and Rosana Lopes. The results in Fig. 4.4 were provided by Dr. Chao for comparison purposes. The analyses presented in Chapter 3 was motivated by the wealth of available literature on Expectation-Maximization.

All other work conducted for the thesis was completed by the student independently.

Funding Sources

Graduate study was supported by funds from National Science Foundation.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

1. INTRODUCTION AND LITERATURE REVIEW

1.1 Cancer Heterogeneity

Cancer development occurs when a normal cell evolves to develop some kind of escape from normal cellular growth control which provides it with a selective growth advantage [3] [4]. Repeated division of such a cell forms tumors that can invade nearby tissues and even spawn metastases. Moreover, the cancerous cells are more likely to develop further mutations compared to normal cells leading to the development of different types of cancer subpopulations within the same tissue [3]. This phenomenon, termed cancer heterogeneity, renders the treatment by a single drug generally ineffective. A "cocktail approach" in cancer drug design could benefit from an estimation of the relative prevalence of the different subpopulations. Information about the relative ratio of the different subpopulations when observed over time also can provide a quantitative measure to compare the effectiveness of different drug combinations on any given cancerous tissue [2]. Hence, modelling cancer heterogeneity holds promise for designing more effective drug combinations.

1.2 Maximum Likelihood and Expectation-Maximization

Maximum likelihood (ML) estimation refers to estimating the model parameters for which the observed data are most likely. An obvious approach to calculate the ML estimate is to set the partial derivatives of the likelihood with respect to the parameters to zero. However, in the case of mixture models, this approach is not well-posed typically because the likelihood function involves the sum of variables inside the logarithmic function. To overcome this problem, numerical algorithms like Newton-Raphson method or gradient-descent can be applied but it usually requires calculation of first and second order derivatives of the likelihood function. This is not possible when the model has hidden variables. Even, in the case of models without hidden variables, Expectation-Maximization [5],[6]

provides a simpler, more robust and easy to implement approach to estimating the maximum likelihood value. Here robustness is considered with respect to missing [7] and/or censored [8],[9] data values. Expectation-Maximization based imputation techniques are quite effective if missing data is small (usually $< 5\%$).

Each iteration of the EM algorithm consists of two steps: the E-step and the M-step. The Expectation or E-step calculates the expectation of complete log-likelihood given the observations and current estimate of parameters. The Maximization or M-step computes the parameter estimates for the next E-step, by maximizing the expectation computed in current E-step with respect to the model parameters. By using the Kullback-Leibler divergence as a measure of information gain or entropy, it can be shown that the log-likelihood never decreases after a combined E-step and M-step [10] and thus convergence is guaranteed.

In this work, a mixture model based on Expectation-Maximization is proposed to estimate the relative ratio of different cancer subpopulations. Each subpopulation is assumed to have a multivariate Gaussian distribution over the observables so that the mixture model is represented by a sum of Gaussians. This assumption provides clean, closed form expressions for the E-step and M-step computations [6] leading to performance gains when compared to Markov Chain Monte-Carlo based techniques. Additionally, being a model-based learning technique, unlike dynamic Bayesian approach proposed in [2], EM does not require information about the response of individual subpopulations or the initial mixture composition while observing the dynamic response (i.e. over time) of the mixture.

One of the limitations of the Expectation-Maximization is the problem of identifiability [11]. When run over multiple iterations, it is not guaranteed that the ratios will always be calculated in a particular order. Even if they do, the algorithm by itself cannot identify the subpopulation corresponding to the estimated parameters. Additionally, although the convergence is guaranteed, the algorithm might converge to local maxima or saddle points

when applied to real data, as discussed in [9] leading to inaccurate estimate. We exploit prior pathway knowledge to overcome these limitations of the EM algorithm.

1.3 Pathway Knowledge

Over a long period of time biologists have collected a wealth of information regarding marginal regulatory interactions within a cell, called pathway knowledge. Unfortunately, this information usually gets ignored in most regulatory network models. In a recent work [12], the authors exploited marginal pathway knowledge to systematically generate Boolean networks using Karnaugh maps. In this work, we use this pathway information to sort the ratios in a fixed order and to choose the correct result from a set of converging points over multiple iterations of the EM algorithm.

2. GAUSSIAN MIXTURE MODEL FOR CANCER HETEROGENEITY

2.1 Model Description

We consider a Boolean model of the mitogen-activated protein kinase (MAPK) signal transduction network proposed in [1] where cancer is represented as a stuck-at fault in the network as shown in Fig. 2.1.

Consider a hypothetical cancer with three major subpopulations with corresponding stuck-at faults locations shown in Table 2.1. Let the mixture be exposed to different drug combinations as shown in Table 2.2. We consider transcription factors FOS-JUN and SP1 as observables of interest for the experiments. We assume that only one transcription factor is observed for each drug combination. If more than one transcription factor is observed, we classify it as separate experiments. As will be evident in later sections, each experiment defines a unique dimension in our multivariate model. Thus, for our classification purposes any unique (drug combination, observable) pair is considered to be a different experiment.

To measure the activity of a transcription factor we measure the expression of a gene activated by that transcription factor. For example, cMYC, JUN and BIRC5 are some genes activated by the transcription factor SP1. We observe the expression of a single gene for the transcription factor in any given experiment. The reporter genes for each

Table 2.1: Stuck-at fault locations for different subpopulations.

Subpopulation	**Type of Fault**	**Location**
Sub. I	Stuck-at one	ERK1/2
Sub. II	Stuck-at one	ERBB2/3
	Stuck-at one	Raf
Sub. III	Stuck-at zero	PTEN

Table 2.2: Drug combinations and reporter genes considered for experiments.

Experiment	**Exp. I**	**Exp. II**	**Exp. III**	**Exp. IV**
Drug Combination	AG1024 + Lapatinib	LY294002 + U0126	U0126	No Drugs
Obs. Transcrip. Factor	SP1	FOS-JUN	SP1	SP1
Reporter Gene	cMYC	CRE31	JUN	BIRC5

Table 2.3: Binary gene response to drug experiments in Table 2.2 for different subpopulations.

Subpopulation	**Exp. I**	**Exp. II**	**Exp. III**	**Exp. IV**
Sub. I	upreg.	upreg.	upreg.	upreg.
Sub. II	upreg.	downreg.	downreg.	upreg.
Sub. III	downreg.	upreg.	downreg.	upreg.

experiment are listed in the last row of Table 2.2. For each experiment, the binary gene response of the corresponding observables can be evaluated using the MAPK network shown in Fig. 2.1. For example, for the first experiment, SP1 is upregulated for the first and second subpopulations while it is downregulated for the third subpopulation. The binary responses of the corresponding observables for all the experiments are summarized in Table 2.3.

In [1], the normalized gene expression ratio is used to measure the activity of the reported genes, which was modeled as the ratio of two normal random variables, each with its standard deviation directly proportional to its mean. This modelling, however, requires the observables to have non-standard prior distribution which leads to non-closed form expressions for the posterior distribution of the relative ratio parameters. To approximate these posterior distributions, variational or Markov-chain Monte Carlo (MCMC) methods

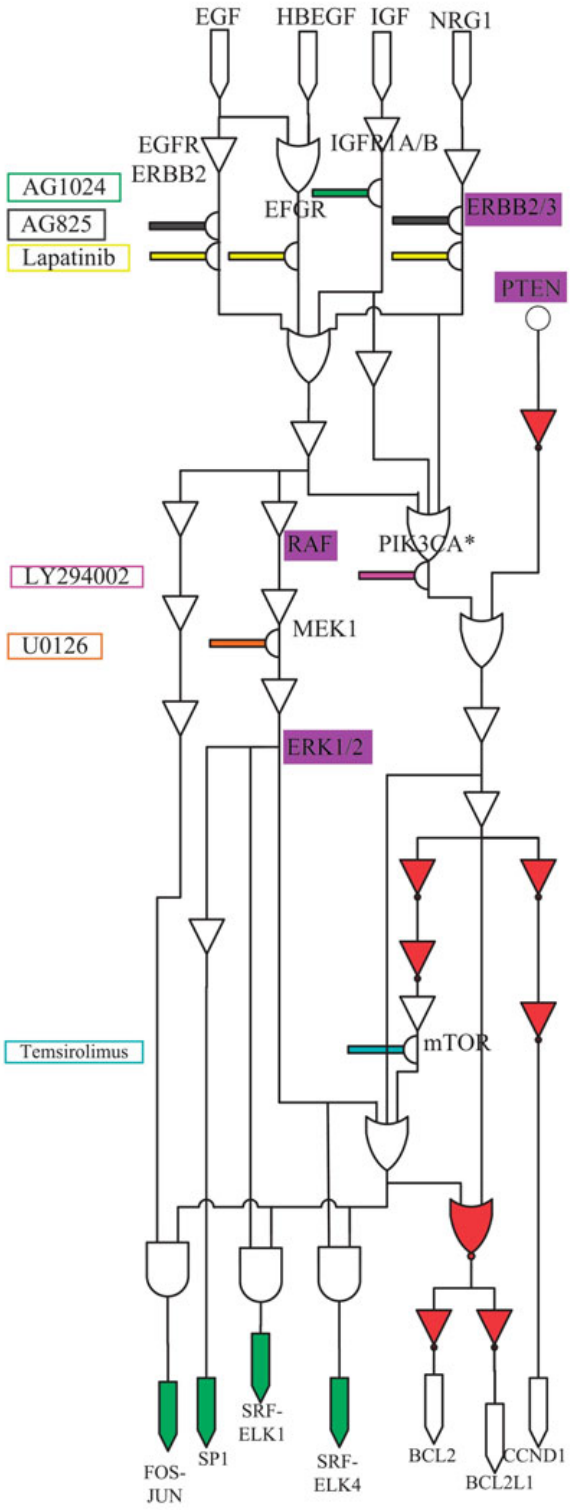

Figure 2.1: MAPK network with stuck-at faults reproduced from Anwoy [1].

are used which are prone to error because of the inherent approximations involved.

In this work, we consider non-normalized gene expression ratio observations. We assume that the gene expression is measured as the intensity of light emitted by a specific fluorescent protein whose production is directly proportional to the expression of the gene to be observed. Since the gene expression is normally distributed with standard deviation directly proportional to its mean [13] we assume that the normal distribution for the observed intensity also exhibits this property. This assumption of a normal distribution facilitates compact vector expression for the posterior distribution and the use of a closed form expression for parameter values at each iteration of the Expectation-Maximization algorithm.

We assume that the intensity of the gene activity in a given subpopulation affected by a drug becomes exactly zero. In reality, however, there might be some feedback network that prevents the gene expression from becoming exactly zero. Even in those cases, it is safe to assume that such a subpopulation will have gene expression values significantly different from that without drug treatment. It is also interesting to note that the work here differs from that in [2] where binary decisions for observed intensity are used in the algorithm. Also, unlike [2], observed intensity for individual subpopulations in response to the drugs is not assumed to be known. Only prior pathway knowledge and drugs causing differential response of given genes are assumed to be known. This information is used to design experiments but the algorithm uses continuous intensity values for improved accuracy.

2.2 Gaussian Mixture Model for Cancer Heterogeneity

Assume $I_{j,k}$ as the unknown mean and $c \times I_{j,k}$ as the standard deviation of the observed intensity distribution for the kth subpopulation in the jth experiment, where the coefficient of variation c is considered to be an unknown parameter. Then, with the assumptions as described above, the observed intensity x_1 for the first experiment has the distribution

defined by the weighted mixture density,

$$p(\mathrm{x}_1) = \alpha_1 \mathcal{N}(\mathrm{x}_1 | I_{1,1}, cI_{1,1}) + \alpha_2 \mathcal{N}(\mathrm{x}_1 | I_{1,2}, cI_{1,2}) \tag{2.1}$$

where $\alpha_1 : \alpha_2 : \alpha_3$ denotes the relative ratio of three subpopulations in the mixture with the constraint

$$\alpha_1 + \alpha_2 + \alpha_3 = 1 \tag{2.2}$$

Similarly, in the second experiment, FOS-JUN is upregulated in the case of the first and third subpopulations and downregulated in the case of the second. Therefore, the probability distribution for the observed intensity x_2 is given by,

$$p(\mathrm{x}_2) = \alpha_1 \mathcal{N}(\mathrm{x}_2 | I_{2,1}, cI_{2,1}) + \alpha_3 \mathcal{N}(\mathrm{x}_2 | I_{2,3}, cI_{2,3}) \tag{2.3}$$

The expressions for the other experiments can be obtained in a similar fashion. This example can be extended to the case when the intensity of the kth subpopulation affected by the drug is not exactly zero by adding the term $\alpha_k \mathcal{N}(\mathrm{x} | I'_{j,k}, c \times I'_{j,k})$ to the mixture density in the jth experiment. Here $I'_{j,k}$ is assumed to be a non-zero mean intensity in the presence of the affecting drug. For compact representation, we introduce the variable $\mu_{k,j}$ for the subpopulation k defined as

$$\mu_{j,k} = d_{j,k} I_{j,k} + (1 - d_{j,k}) I'_{j,k} \tag{2.4}$$

where $d_{j,k}$ is a binary value which equals '1' if the drug in experiment j does not affect the observable in subpopulation k, and otherwise the value equals '0'. The observed intensity

for the jth experiment can now be expressed in the general form as,

$$p(\mathrm{x}_j) = \sum_{k=1}^{3} \alpha_k \mathcal{N}(\mathrm{x}_j | \mu_{j,k}, c\mu_{j,k}) \tag{2.5}$$

To combine information across all experiments, we can further compact Eq. (2.5) to vector representation in terms of the multivariate Gaussian distribution as

$$p(\mathbf{x}) = \sum_{k=1}^{3} \alpha_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{2.6}$$

where $\mathbf{x}$ is the column vector $[\mathrm{x}_1, \mathrm{x}_2, \mathrm{x}_3, \mathrm{x}_4]^T$ and $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is a multivariate Gaussian distribution defined as

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{4/2}|\boldsymbol{\Sigma}_k|^{1/2}} \times \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T {\boldsymbol{\Sigma}_k}^{-1} (\mathbf{x} - \boldsymbol{\mu}_k)\right) \tag{2.7}$$

$\boldsymbol{\mu}_k$ is a column vector of means $[\mu_{1,k}, \mu_{2,k}, \mu_{3,k}, \mu_{4,k}]^T$ for the kth subpopulation given by

$$\boldsymbol{\mu}_k = diag\left(\begin{bmatrix} \mathbf{d}_k & (1_{4\times 1} - \mathbf{d}_k) \end{bmatrix} \begin{bmatrix} \mathbf{I}_k \\ \mathbf{I'}_k \end{bmatrix}\right) \tag{2.8}$$

where $\mathbf{d}_k$ is the 4×1 binary-valued "expression profile" vector with jth entry $d_{j,k}$ = 1 if corresponding transcription factor is upregulated in jth experiment and zero otherwise. Here we have assumed that the drugs are kinase-inhibitors which downregulate the affected transcription factor. $1_{4\times 1}$ is column vector with all entries as 1. $\mathbf{I}_k$ and $\mathbf{I'}_k$ are 1×4 column vectors of mean intensity when the transcription factor is upregulated and downregulated respectively. Eq. (2.8) is just matrix based notation of Eq. (2.4). The variable $\boldsymbol{\Sigma}_k$ is introduced as a variance-covariance matrix for the kth subpopulation.

The unknown parameters in this model are the weight vector $\boldsymbol{\alpha}$, mean intensity matrix

$\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$. These parameters need to be estimated from observed intensity values $\mathbf{x}$. Expectation-Maximization iteratively updates the parameters in the direction of increasing likelihood of observed intensity.

3. EXPECTATION-MAXIMIZATION FOR CANCER HETEROGENEITY

3.1 Mixture Model Representation with Hidden Variables

In this chapter we consider a more generalized case of the example presented in Chapter 2. Consider an ensemble of K subpopulations with J unique (drug combination, observable) pairs, henceforth referred to as J experiments for simplicity. Let the weight vector $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_K]$ denote the relative ratio of different subpopulations with $\sum_{k=1}^{K} \alpha_k = 1$ and $\alpha_k > 0$. From Eq. (2.6), the probability distribution of the ith data point is dependent on the weight vector $\boldsymbol{\alpha}$, the mean intensity matrix $\boldsymbol{\mu}$, the three-dimensional covariance matrix $\boldsymbol{\Sigma}$ and the expression profile matrix $\mathbf{d}$ =[$\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_k$]. The conditional probability distribution of the ith observed data point $\mathbf{x}_i \in \mathbb{R}^J$ is given by

$$p(\mathbf{x}_i|\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{d}) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \mathbf{d}) \tag{3.1}$$

Expectation-Maximization is an iterative procedure to maximize the log-likelihood of above expression. To make the estimation of parameters more tractable, we introduce hidden data z_i to denote hidden assignment of data point $\mathbf{x}_i$ to one of the subpopulations. In other words, $z_i = k$ means ith observed data point was sampled from kth subpopulation. This information is not directly observed and hence the term "hidden". Assume z_i are modeled as discrete random variables $Z_i \in \{1, 2, ..., K\}$, i.e. z_i is a realization of random variable Z_i. Assuming $\mathbf{x}_i$ as a realization of the continuous random variable $\mathbf{X}_i$, the joint probability over the "complete data set" $\mathbf{Y}_i = (\mathbf{X}_i, Z_i)$ including both observed and hidden data is given by

$$p(\mathbf{y}_i|\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{d}) = p(\mathbf{x}_i, z_i|\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{d}) = P(z_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{d}) p(\mathbf{x}_i|z_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{d}) \tag{3.2}$$

Since $p(\mathbf{x}_i|z_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{d})$ is assumed to be normally distributed with mean $\boldsymbol{\mu}_{z_i}$ and standard deviation $\boldsymbol{\Sigma}_{z_i}$, we have

$$p(\mathbf{y}_i|\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{d}) = P(z_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{d})\mathcal{N}(\mathbf{x}_i|z_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{d}) \tag{3.3}$$

where $\mathcal{N}(\mathbf{x}_i|Z_i = k, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{d}) = \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \mathbf{d})$. Using the theorem of total probability the probability distribution of the ith data point is given by marginalizing Eq. (3.3) over the domain of Z_i as

$$p(\mathbf{x}_i|\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{d}) = \sum_{z_i=1}^{K} P(z_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{d})\mathcal{N}(\mathbf{x}_i|z_i, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{d}) \tag{3.4}$$

Note that since Z_i is a discrete random variable which denotes which subpopulation each data point came from, we have $P(Z_i = k|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{d}) = \alpha_k$ which verifies that Eq. (3.1) and (3.4) are equal. However, the benefit of using the latter is that it simplifies the log-likelihood expression.

3.2 Expectation-Maximization

Expectation-Maximization is an iterative method that tries to maximize the expected log-likelihood of the complete data $\mathbf{Y}_i$ using the following iterative procedure: first it makes a guess about the complete data $\mathbf{Y}_i$, then it solves for parameters that maximize (the expected) log-likelihood of $\mathbf{Y}_i$. Once an estimate for parameters is obtained, it is then used to make a better guess of the complete data $\mathbf{Y}_i$, and iterations continue till some threshold is reached. In the following sections, E-step and M-step are discussed in detail.

3.2.1 E-step

In the E-step, the expected value of the log-likelihood of $\mathbf{Y}_i$ is calculated from the conditional probability distribution of $\mathbf{Y}_i$ given the observation $\mathbf{x}_i$ and the current estimate

of parameters. For compactness in representation, let $\boldsymbol{\theta} = \{\boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{d}\}$ and let $\boldsymbol{\theta}^{(t)}$ denote the estimated parameters at the end of the tth iteration. Then,

$$E_{\mathbf{Y}_i|\mathbf{x}_i,\boldsymbol{\theta}^{(t)}}[\log p(\mathbf{Y}_i|\boldsymbol{\theta})] = \int_{\mathcal{Y}(\mathbf{x}_i)} \log[p(\mathbf{y}_i|\boldsymbol{\theta})] p(\mathbf{y}_i|\mathbf{x}_i, \boldsymbol{\theta}^{(t)}) d\mathbf{y}_i \tag{3.5}$$

where $\mathcal{Y}(\mathbf{x}_i)$ is the support of $\mathbf{Y}_i$ given $\mathbf{x}_i$, i.e. closure of the set $\{\mathbf{y}_i | p(\mathbf{y}_i|\mathbf{x}_i) > 0)\}$. Since the only random part in the complete data $\mathbf{Y}_i$ is Z_i the expectation of $p(\mathbf{Y}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{d})$ can be calculated over the domain of Z,

$$\begin{aligned}
E_{\mathbf{Y}_i|\mathbf{x}_i,\boldsymbol{\theta}^{(t)}}[\log p(\mathbf{Y}_i|\boldsymbol{\theta})] &= E_{Z_i|\mathbf{x}_i,\boldsymbol{\theta}^{(t)}}[\log p(\mathbf{x}_i, Z_i|\boldsymbol{\theta})] \\
&= \sum_{z_i=1}^{K} P(z_i|\mathbf{x}_i, \boldsymbol{\theta}^{(t)}) \log p(\mathbf{x}_i, z_i|\boldsymbol{\theta}) \\
&= \sum_{z_i=1}^{K} \frac{p(z_i, \mathbf{x}_i|\boldsymbol{\theta}^{(t)})}{p(\mathbf{x}_i|\boldsymbol{\theta}^{(t)})} \log p(\mathbf{x}_i, z_i|\boldsymbol{\theta}) \\
&= \sum_{z_i=1}^{K} \frac{p(\mathbf{y}_i|\boldsymbol{\theta}^{(t)})}{p(\mathbf{x}_i|\boldsymbol{\theta}^{(t)})} \log p(\mathbf{x}_i, z_i|\boldsymbol{\theta}) \\
&= \sum_{k=1}^{K} \gamma_{ik}^{(t)} \log(\alpha_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \mathbf{d})) \\
&= \sum_{k=1}^{K} \gamma_{ik}^{(t)} \left(\log \alpha_k - \frac{1}{2} \log |\boldsymbol{\Sigma}_k| - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right)
\end{aligned} \tag{3.6}$$

where we have ignored the constants in last step and $\gamma_{ik}^{(t)}$ is defined as

$$\gamma_{ik}^{(t)} \triangleq \frac{p(z_i, \mathbf{x}_i|\boldsymbol{\theta}^{(t)})}{p(\mathbf{x}_i|\boldsymbol{\theta}^{(t)})} = \frac{\alpha_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \mathbf{d})}{\sum_{k=1}^{K} \alpha_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \mathbf{d})} \tag{3.7}$$

Let $\mathbf{x}$ denote the collection of all N observations $\mathbf{x_i}$ taken together modelled as continuous random variable $\mathbf{X}$. Let $\mathbf{Y} = (\mathbf{X}, Z)$ where Z is collection of hidden random variables Z_i for all observations. Then, assuming all data points to be i.i.d., it can be shown that the

expected log-likelihood of $p(\mathbf{Y}|\boldsymbol{\theta})$ is given by

$$E_{\mathbf{Y}|\mathbf{x},\boldsymbol{\theta}^{(t)}}[\log p(\mathbf{Y}|\boldsymbol{\theta})] = \sum_{i=1}^{N} E_{\mathbf{Y}_i|\mathbf{x}_i,\boldsymbol{\theta}^{(t)}}[\log p(\mathbf{Y}_i|\boldsymbol{\theta})] \tag{3.8}$$

Substituting Eq. (3.6) in (3.8)

$$E_{\mathbf{Y}|\mathbf{x},\boldsymbol{\theta}^{(t)}}[\log p(\mathbf{Y}|\boldsymbol{\theta})] = \sum_{i=1}^{N}\sum_{k=1}^{K} \gamma_{ik}^{(t)} \left(\log \alpha_k - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k) \right) \tag{3.9}$$

3.2.2 M-step

The M-step is to find the parameters that maximize the expectation computed in E-step, i.e.,

$$\begin{aligned} \boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} & E_{\mathbf{Y}|\mathbf{x},\boldsymbol{\theta}^{(t)}}[\log p(\mathbf{Y}|\boldsymbol{\theta})] \\ & \text{subject to} \\ & \sum_{k=1}^{K} \alpha_k = 1, \\ & \alpha_k \geq 0, k = 1, 2, ..., K, \\ & \Sigma_k \succ 0, k = 1, 2, ..., K. \end{aligned} \tag{3.10}$$

where $\Sigma_k \succ 0$ means Σ_k is positive definite. To find the maximizing parameters, we need to set partial derivative with respect to each parameter to zero. For ease in partial differentiation, Eq. (3.9) can be expressed as

$$\begin{aligned} E_{\mathbf{Y}|\mathbf{x},\boldsymbol{\theta}^{(t)}}[\log p(\mathbf{Y}|\boldsymbol{\theta})] = & \sum_{k=1}^{K} \left(\sum_{i=1}^{N} \gamma_{ik}^{(t)} \right) \left(\log \alpha_k - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| \right) \\ & - \sum_{i=1}^{N}\sum_{k=1}^{K} \frac{1}{2} \gamma_{ik}^{(t)} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k) \end{aligned} \tag{3.11}$$

Setting partial derivatives of expression in Eq. (3.11) to zero we get

$$\alpha_k^{(t+1)} = \frac{\sum_{i=1}^{N} \gamma_{ik}^{(t)}}{N} \tag{3.12}$$

$$\boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{i=1}^{N} \gamma_{ik}^{(t)} \mathbf{x}_i}{\sum_{i=1}^{N} \gamma_{ik}^{(t)}} \tag{3.13}$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_{i=1}^{N} \gamma_{ik}^{(t)} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T}{\sum_{i=1}^{N} \gamma_{ik}^{(t)}} \tag{3.14}$$

3.3 Initialization, Convergence and Exploiting Pathway Knowledge

An important limitation of mixture models in general is the problem of identifiability [11], i.e., in a model with K-densities there are $K!$ possible ways of assigning K parameters to K cancer subpopulations. In case of cancer, the interchange of relative ratios of different subpopulations can be lethal. This problem is addressed by exploiting the prior pathway knowledge. By comparing the observed mean intensities with the expression profile vector of each subpopulation, the parameters can be corrrectly matched to corresponding subpopulations. Moreover, it is important to note that although EM tries to find peaks in log-likelihood $L(\boldsymbol{\theta}|\mathbf{x})$, if the latter has multiple peaks then the algorithm might converge to local maxima. The convergence is quite sensitive to initializations and solution might converge to maxima closest to the initial point. To overcome this, we run the algorithm over multiple initializations and choose correct result by defining an error metric based on pathway knowledge.

In the next chapter, we present results of estimation of cancer heterogeneity on synthetic and real experiment data. In addition, we employ this algorithm to observe dynamic behaviour (over time) of mixture of cancer subpopulations. The results are compared with the case when complete information about individual subpopulations and initial mixture distribution is known [2].

4. RESULTS

4.1 Synthetic Data

To validate the algorithm we generated synthetic data for the scenario discussed in Chapter 2. We assume that when an observable is downregulated, the corresponding observed intensity drops to one fourth of the upregulated intensity. The value of the coefficient of variation c is taken to be equal to 0.17 and assumed constant for all cases. Experimental studies have verified that typical values of c are close to this [13]. The assumed mean intensity values for upregulated observables for different experiments are given in Table 4.1.

Since the intensity values cannot be negative, we set the values generated from normal distributions around these means to 1000 whenever they are negative. The data was generated for a total of 100 cells for each experiment with 20 cells from subpopulation I, 30 cells from subpopulation II and 50 cells from subpopulation III. The algorithm is repeated for 1000 times with different random initial values every time. The calculated ratio is recorded for each repeat and kernel densities were plotted for the recorded ratios for the mixture as shown in Fig. 4.1. Since there is a single peak in density, the algorithm converged to global maxima. The ratio corresponding to the peak was found to be $(0.20, 0.30, 0.50)$.

4.2 Experimental Data

A mixture of three cancer cell lines, A2068, HCT116 and SW480 was prepared in the ratio $(0.15, 0.35, 0.50)$. The intensity of light was observed over red, blue and green

Table 4.1: Assumed intensity values for generating synthetic data.

Observables	**cMYC**	**CRE31**	**JUN**	**BIRC5**
Intensity	1×10^3	5×10^3	1.3×10^3	7.6×10^3

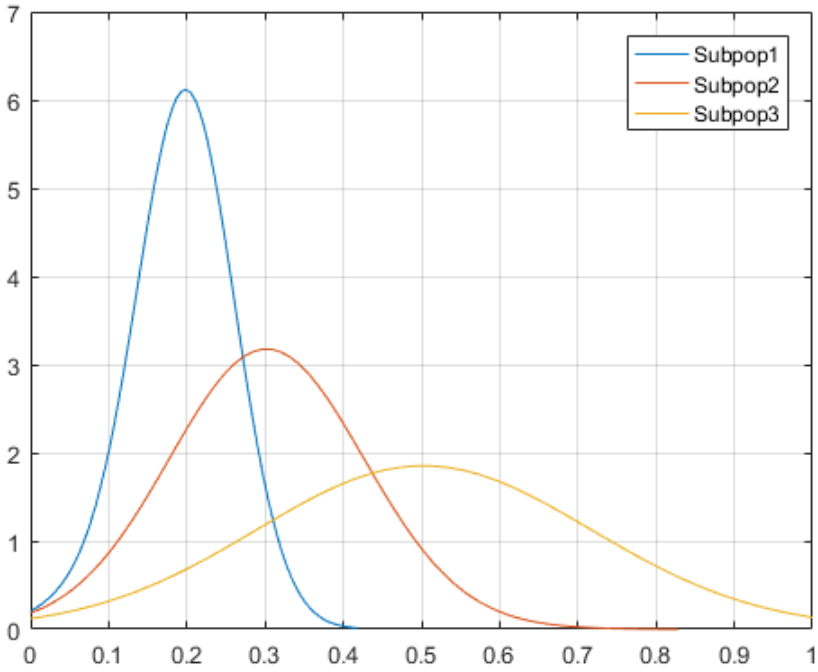

Figure 4.1: Kernel density estimate of relative ratio parameters for synthetic data.

Table 4.2: Error values of mean intensities for two peaks in Fig. 4.2.

T	$\sum_{j \in \text{downreg genes}} \lvert I_{j1T} - I_{j1}^{avg} \rvert$	$\sum_{j \in \text{downreg genes}} \lvert I_{j2T} - I_{j2}^{avg} \rvert$
1	1142	23.9
2	861	8
3	9	356
E_p	**2012**	**387.9**

channels for a group of cells drawn randomly from the mixture. The prior information provided about each cell line is shown in Table 4.3. The distribution of the calculated ratios over multiple runs of the EM algorithm are shown in Fig. 4.2. It is evident from the figure that the EM algorithm converges to two local maximas. To determine, the accurate composition from these two choices, we can exploit the pathway knowledge. Since the mixture is untreated, we do not expect sudden changes in the initial hours. Data observed at time T = 1, T = 2 and T = 3 hours was combined and the resultant mean intensities from the algorithm were compared against those obtained from each time interval separately. We defined the error metric E_p of peak p as,

$$E_p = \sum_{T=1}^{3} \sum_{j \in \text{downreg genes}} |I_{jpT} - I_{jp}^{avg}| \tag{4.1}$$

where I^{avg} is the intensity from combined data for initial three hours. Here we chose downregulated genes only because mean intensity and variance and hence the error would be less for downregulated genes compared to upregulated genes. The peak with the minimum error E_p is chosen as the final result. From E_p values in Table 4.2 it second peak is the correct choice. The calculated mean intensities in the second case show least variation over first three hours.

The mean intensity estimated by the EM algorithm and the corresponding ratios are

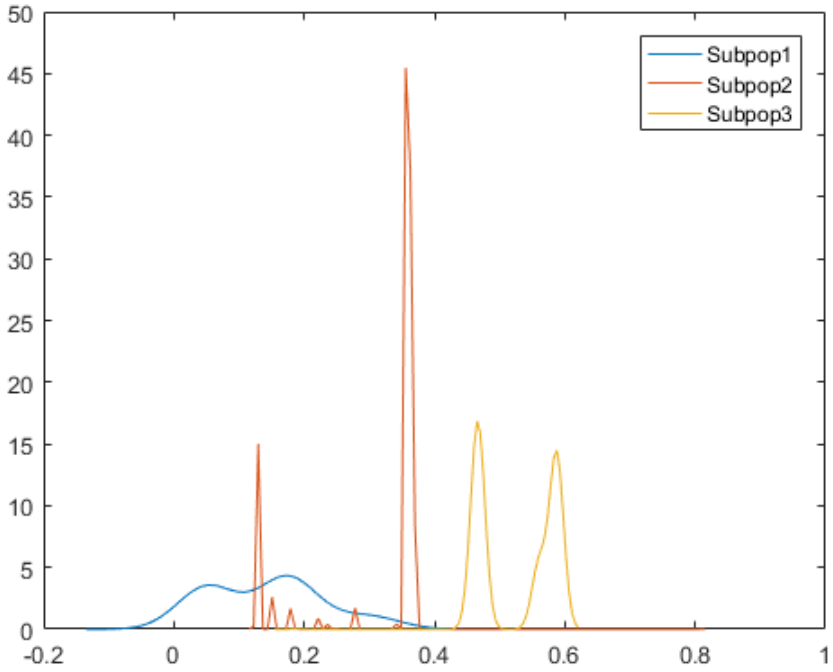

Figure 4.2: Kernel density estimate of relative ratio parameters for real data at time T = 1.

Table 4.3: Prior pathway knowledge for experimental data in Section 4.2.

Subpopulation	**Red Channel**	**Green Channel**	**Blue Channel**
A2058	downreg.	upreg.	upreg.
HCT116	upreg	downreg.	upreg.
SW480	downreg.	downreg.	upreg.

Table 4.4: Calculated and observed mean intensity values for different observables.

Subpop.	**Channel**	**Calculated Mean**	**Calculated ratio**	**Observed Mean**
Subpop. I	R	4.7602	0.1785	3.5703
	G	1.189×10^3		1.0576×10^3
	B	9.173×10^3		7.0516×10^3
Subpop. II	R	2.477×10^3	0.3531	2.1739×10^3
	G	6.0692		3.6386
	B	1.0247×10^4		8.1546×10^3
Subpop. III	R	0.000	0.4684	5.8139
	G	4.3		4.2163
	B	9.3755×10^3		8.7397×10^3

shown in Table 4.4. Comparing the observed mean intensities with pathway knowledge, we can conclude that subpopulation I, II and III are A2058, HCT116 and SW480 respectively. Thus, we can conclude that the ratio of cell-lines (A2068, HCT116, SW480) is $(0.1785, 0.3531, 0.4684)$. Also, in Table 4.4 observed mean intensity values of the cell-lines observed individually are shown for comparison with the calculated ones.

Furthermore, to observe the change in heterogeneity with time, the results of algorithm at different time instants are shown in Fig. 4.3. To break the tie in cases where multiple peaks were obtained we used intensities I_{j2}^{avg} as metric for comparison, provided in Table 4.5. The variation of calculated intensities of downregulated variables from the standard

Table 4.5: Calculated average mean intensity values for combined T = 1, 2 and 3.

Subpop.	**Channel**	**Intensity** ($\times 10^4$)
Subpop. I	R	0.0051
	G	0.1243
	B	1.0933
Subpop. II	R	0.2356
	G	0.0004
	B	1.2383
Subpop. III	R	0.000
	G	0.0004
	B	1.1669

in Table 4.5 and corresponding ratios at different time instants are given in Table 4.6. The results of a recently proposed algorithm [2] when applied on our data are shown in Fig. 4.4. This result was provided by Dr. Chao for comparison purposes. Unlike this work, the dynamic Bayesian framework based algorithm proposed in [2] assumes that complete information about each individual subpopulation is known. In other words, the distribution of intensities and initial mixture proportion are assumed to be known. Comparing Fig. 4.3 and 4.4, it can be seen that even without this information our results closely estimate the dynamic behaviour predicted with complete information.

Table 4.6: Variation of mean intensities from standard 4.5 for data observed over time.

T	Calculated Ratio	Variation of calculated intensities
1	$(0.1730, 0.3607, 0.4663)$	23.9
5	$(0.1628, 0.3775, 0.4596)$	15
10	$(0.1785, 0.3792, 0.4422)$	46
15	$(0.1617, 0.3996, 0.4387)$	35
20	$(0.1639, 0.4199, 0.4162)$	16
25	$(0.1666, 0.4425, 0.3908)$	19
30	$(0.1834, 0.4520, 0.3647)$	143
35	$(0.1956, 0.4627, 0.3417)$	205
40	$(0.2087, 0.4760, 0.3153)$	216
45	$(0.2176, 0.4883, 0.2941)$	287.7

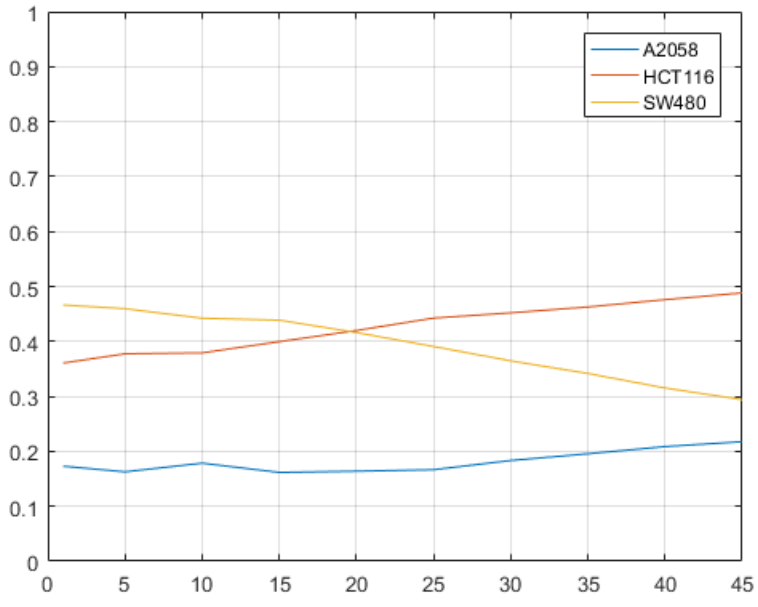

Figure 4.3: Estimate of relative ratio parameters over time.

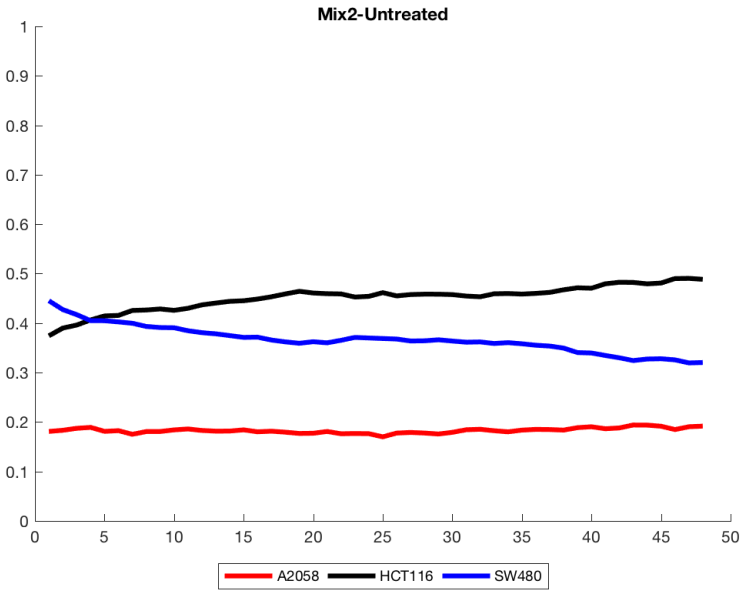

Figure 4.4: Estimate of relative ratio parameters over time using algorithm in [2].

5. SUMMARY AND CONCLUSIONS

A Gaussian Mixture-Model with Expectation-Maximization was proposed for addressing cancer heterogeneity. The working of the algorithm was verified on synthetic and experimental data. Future work may involve verification on datasets involving drug-treated mixtures and comparing efficacy of different drugs in killing heterogeneous cancer cell populations.

REFERENCES

[1] A. K. Mohanty, A. Datta, and V. Venkatraj, "A model for cancer tissue heterogeneity," *IEEE Transactions on Biomedical Engineering*, vol. 61, no. 3, pp. 966–974, 2014.

[2] C. Sima, J. Hua, R. Lopes, A. Datta, and M. L. Bittner, "Detecting cell growth and drug response in heterogeneous populations: A dynamic imaging approach," in *Bioinformatics and Bioengineering (BIBE), 2016 IEEE 16th International Conference on*, pp. 121–128, IEEE, 2016.

[3] P. C. Nowell, "The clonal evolution of tumor cell populations," *Science*, vol. 194, no. 4260, pp. 23–28, 1976.

[4] R. Weinberg, *The Biology of Cancer*. New York: Garland Science, 2013.

[5] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (methodological)*, pp. 1–38, 1977.

[6] R. M. Neal and G. E. Hinton, "A view of the em algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*, pp. 355–368, Springer, 1998.

[7] S. L. Lauritzen, "The em algorithm for graphical association models with missing data," *Computational Statistics & Data Analysis*, vol. 19, no. 2, pp. 191–201, 1995.

[8] H. Ng, P. Chan, and N. Balakrishnan, "Estimation of parameters from progressively censored data using em algorithm," *Computational Statistics & Data Analysis*, vol. 39, no. 4, pp. 371–386, 2002.

[9] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, vol. 382. John Wiley & Sons, 2007.

[10] C. J. Wu, "On the convergence properties of the em algorithm," *The Annals of Statistics*, pp. 95–103, 1983.

[11] G. Casella and R. L. Berger, *Statistical Inference*, vol. 2. Pacific Grove, CA: Duxbury Press, 2002.

[12] R. K. Layek, A. Datta, and E. R. Dougherty, "From biological pathways to regulatory networks," *Molecular BioSystems*, vol. 7, no. 3, pp. 843–851, 2011.

[13] Y. Chen, E. R. Dougherty, and M. L. Bittner, "Ratio-based decisions and the quantitative analysis of cdna microarray images," *Journal of Biomedical Optics*, vol. 2, no. 4, pp. 364–374, 1997.