Supervised clustering of high dimensional data using regularized mixture modeling

Wennan Chang
Department of Electrical and
Computer Engineering
Purdue University

Changlin Wan
Department of Electrical and
Computer Engineering
Purdue University

Yong Zang
Department of Biostatistics
Indiana University

Chi Zhang*
Medical and Molecular Genetics
Indiana University School of Medicine
Email: czhang87@iu.edu

Sha Cao*
Department of Biostatistics
Indiana University
Email: shacao@iu.edu

***Abstract*—Identifying relationships between molecular variations and their clinical presentations has been challenged by the heterogeneous causes of a disease. It is imperative to unveil the relationship between the high dimensional molecular manifestations and the clinical presentations, while taking into account the possible heterogeneity of the study subjects. We proposed a novel supervised clustering algorithm using penalized mixture regression model, called CSMR, to deal with the challenges in studying the heterogeneous relationships between high dimensional molecular features to a phenotype. The algorithm was adapted from the classification expectation maximization algorithm, which offers a novel supervised solution to the clustering problem, with substantial improvement on both the computational efficiency and biological interpretability. Experimental evaluation on simulated benchmark datasets demonstrated that the CSMR can accurately identify the subspaces on which subset of features are explanatory to the response variables, and it outperformed the baseline methods. Application of CSMR on a drug sensitivity dataset again demonstrated the superior performance of CSMR over the others, where CSMR is powerful in recapitulating the distinct subgroups hidden in the pool of cell lines with regards to their coping mechanisms to different drugs. CSMR represents a big data analysis tool with the potential to resolve the complexity of translating the clinical manifestations of the disease to the real causes underpinning it. We believe that it will bring new understanding to the molecular basis of a disease, and could be of special relevance in the growing field of personalized medicine.**

I. Introduction

Detection and estimation of the molecular markers associated with phenotypic features is one of the most important problems in biomedical research. Predicative models have been extensively used to link molecular markers to a phenotypic trait, however, the unobserved patient heterogeneity obfuscates the effort to build a unified model that works for all hidden disease subtypes. It has been well understood that various subtypes exist for many common diseases, which vary in etiology, pathogenesis, and prognosis [1], [2], [3]. For example, the cancer cells are constantly evolving in the tumor microenvironment, and they may acquire variations on alternative pathways in response to treatment, which explains why certain patients have better prognoses than others in response to the same treatment [4]. This implies that the same predicative model that links molecular markers to a phenotypic trait may not be valid for every patient, and further it is unclear to what extent the patients should be considered together [5]. Therefore, it is judicious to construct a set of heterogeneous models, each of which corresponds to one subtype.

The fast advancement in high-throughput technology has transformed the biomedical research ecosystem by scaling data acquisition, providing us with unprecedented opportunity to interrogate biology in novel and creative ways. For cancer research, the Broad Institute Cancer Cell Line Encyclopedia (CCLE) [6], Cancer Therapy Response Portal (CTRP) v1/v2 [7], and Genomics of Drug Sensitivity in Cancer (GDSC) [8] datasets contain 24, 185, 481, and 261 drug compound screening data for 504, 242, 860, and 1001 cell lines respectively, together with the multi-omic profiles of the cell lines; the Cancer Genome Atlas (TCGA) has collected biospecimens and matched clinical phenotypes for over 10,000 cancer patients [9]. Consequently, for each sample, there is a tremendous amount and variety of data: 20,000 genes expression profiles, 1 million single-nucleotide polymorphism genotypes, exome and whole-genome sequences, methylation of tens of thousands of CpG islands and the expression of microRNA. From this plurality of data, we anticipate that exploratory methods will serve to extract and characterize molecular subgroups relevant to phenotypic outcomes. However, the growing number of variables does not necessarily increase the discriminative power in classification [10]. To identify the most important molecular biomarkers, variable selection is one of the most commonly used approaches. In particular, the penalized regularization methods have received a great deal of attention [11], [12], [13]. Despite major progress in the research of penalized regression, heterogeneity in variable selection of high dimensional feature spaces remains to be challenging.

Unsupervised learning algorithms such as finite mixture models are typically employed to deal with heterogeneity in

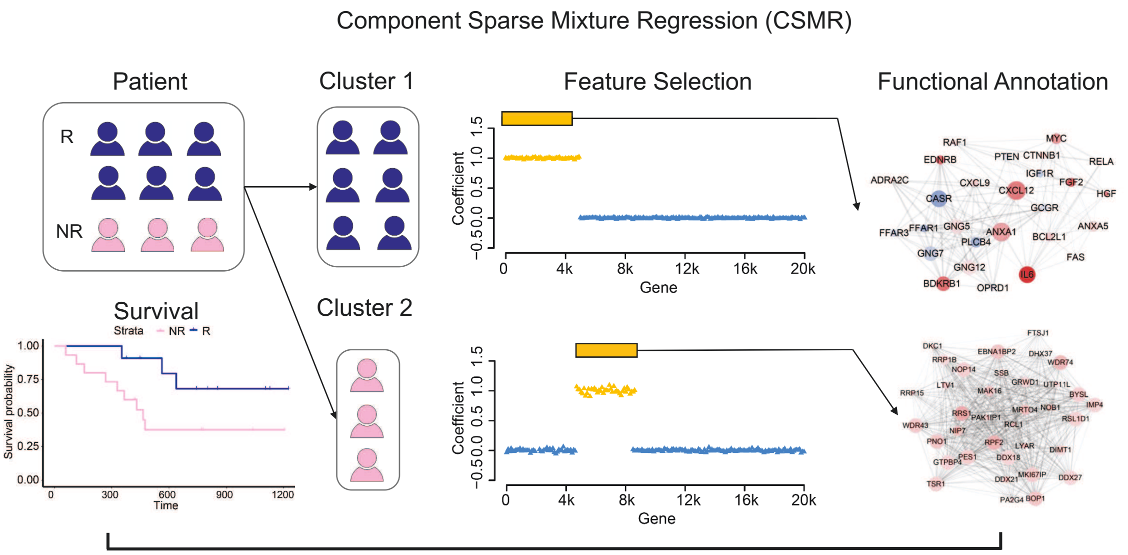

Fig. 1: The motivation of CSMR. Under the same treatment, some patients acquired one mechanism to deal with the drug, (blue), while others picked up another (pink), resulting in different prognoses to the same treatment. The motivation of CSMR is to cluster the patients in a supervised fashion and examine what are the genes (yellow) that are selected in tumor progression that lead to the different drug resistance subtypes of patients, and their functions (network).

subpopulations by assuming a separate distribution for each subpopulation [14], [15], [16]. Based on the molecular subtypes, deeper investigation into the molecular and phenotypic distinctions within each subtype could be carried out. Although the clustering methods may produce satisfactory classification of subtypes, it does not select molecular markers distinctive for each subtype, which however is essential in precision medicine. And more importantly, without any supervision, the defined clusters based on a sea of molecular features may not necessarily relate to the phenotype of interest.

Hence, to study the heterogeneous relations of molecular markers to a certain phenotypic trait, our challenges is distinct in two ways: the variables of interest to each subgroup may be a distinct and sparse set of the high dimensional molecular features, and the set of patients in each subgroup is not known. In this article, we proposed a novel and efficient supervised clustering algorithm based on penalized mixture regression model that synergizes with potential heterogeneity in high dimensional regression problem. Essentially, we assume that observations belong to unlabeled classes with class-specific regression models relating their unique and selective molecular markers to the phenotypic outcome.

II. Preliminaries

Since first introduced in [17], finite mixture Gaussian regression (FMGR) has been extensively studied and widely used in various fields [18], [19], [20], [14], [21], [22]. Let $Y = (y_1, ..., y_N)^T \in \mathcal{R}^N$, $X = [\boldsymbol{x}_1, ..., \boldsymbol{x}_N]^T \in \mathcal{R}^{N\times(P+1)}$ be a finite set of observations, and X the design matrix with intercept and P independent variables, and Y the response vector. Consider an FMGR model parameterized by $\boldsymbol{\theta} = \{(\pi_k, \boldsymbol{\beta}_k, \sigma_k^2)\}_{k=1}^K$, it is assumed that when the i-th observation, $(\boldsymbol{x}_i, y_i)$, belongs to the k-th component, $k = 1, ..., K$, then $y_i = \boldsymbol{x}_i^T\boldsymbol{\beta}_k + \epsilon_{ik}$, and $\epsilon_{ik} \sim N(0, \sigma_k^2)$. In other words, the conditional density of y given $\boldsymbol{x}$ is $f(y|\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{k=1}^K \pi_k \mathcal{N}(y; \boldsymbol{x^T\beta_k}, \sigma_k^2)$, where $\mathcal{N}(y; \mu, \sigma^2)$ is the normal density function with mean μ and variance σ^2. And the log-likelihood for observations $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ is

$$\mathcal{L}(\boldsymbol{\theta}) = \Sigma_{i=1}^N \log(\Sigma_{k=1}^K \pi_k \mathcal{N}(y_i; \boldsymbol{x}_i^T\boldsymbol{\beta}_k, \sigma_k^2)) \tag{1}$$

The unknown parameters $\boldsymbol{\theta}$ can be estimated by the maximum likelihood estimator (MLE), which maximizes (1). Note that the maximizer of (1) does not have an explicit solution and is usually solved by the EM algorithm. Basically, the EM algorithm maximizes the complete log likelihood function, $\mathcal{L}^c(\boldsymbol{\theta})$, through iterative steps, which is defined by

$$\mathcal{L}^c(\boldsymbol{\theta}) = \Sigma_{i=1}^N \Sigma_{k=1}^K z_{ik}[\log \pi_k + \log \mathcal{N}(y_i; \boldsymbol{x}_i^T\boldsymbol{\beta}_k, \sigma_k^2)] \tag{2}$$

where z_{ik} is a cluster indicator variable, and $z_{ik} = 1$ if the i-th observation belongs to the k-th cluster, and 0 otherwise.

While mixture regression model is capable of handling the heterogeneous relationships, it doesn't work in the case of high dimensional molecular features, where the total number of parameters to be estimated is far more than the total number of observations. In addition, with the dense linear coefficients given by the ordinary EM algorithm, it is hard to deduce the disease subtype specific molecular markers and make meaningful interpretations.

Penalized mixture regression has been explored in different settings [23], [24], [25], [26], [27] to handle the high dimensional mixture regression problem. The variable selection problem in finite mixture of regression model was first studied using regularization methods such as LASSO [11] and SCAD [13] in [23], called FMRS. They considered the traditional cases when the number of candidate covariates is much smaller than the sample size, and proposed a modified expectation-maximization (EM) algorithm to perform both estimation and variable selection simultaneously. In [24], the authors proposed a reparameterized mixture of regressions model, and showed evidence for the advantage of multiple components that can be exploited for variable selection over non-mixture linear regression. A block-wise Minorization Maximization (MM) algorithm was proposed in [26], where at each iteration, the likelihood function is maximized with respect to a block of variables while the rest of the blocks are held fixed. To solve the population heterogeneity and feature selection problems, an imputation-conditional consistency (ICC) algorithm was proposed by [27], resulting in consistent estimators. While some of the methods may produce consistent estimates of $\boldsymbol{\theta}$ under proper conditions, they tend to suffer from slow convergence rate in high dimensional setting, especially with smaller N or larger K, and the number of hyper-parameters for regularization further drags down the computational efficiency caused by the need of cross validation. We here propose a novel algorithm based on classification EM for penalized mixture regression to circumvent these existing challenges in clustering high dimensional data using mixture regression, which largely increased the computational efficiency.

The rest of the article is organized as follows: in Section 3, we introduce our algorithm, **C**omponent-wise **S**parse **M**ixture **R**egression (**CSMR**); in section 4, we compare CSMR with four state-of-the-art algorithms on synthetic datasets, in section 5, we applied all the five algorithms on 24 drug sensitivity data in CCLE, to screen for genes that underlie the heterogeneous drug resistance mechanisms.

III. Methods

We assume that the samples belong to different sub-populations, each of which is defined by a distinct relationship between the molecular biomarkers to the phenotype of interest, and the molecular markers are sparse subsets of the high dimensional molecular profiles specific to each sub-population. Figure 1 illustrated an example where the patients fall under two distinct subgroups: blue for patients acquiring one mechanism to the treatment that resulted in responsiveness, while pink for patients acquiring another mechanism to the same drug that resulted in non-responsiveness. The goal of our method is to cluster the samples (blue and pink) supervised by the patients drug sensitivity measure, and find the defining molecular features (yellow) associated with each cluster. The identified molecular features could be further studied to guide targeted therapeutic designs.

A. The penalized likelihood of mixture regression

Knowing that $\boldsymbol{\beta}_k$ is sparse means many elements in $\boldsymbol{\beta}_k$ will tend to be close to zero, but not exactly zero without proper regularization in the model. To simultaneously shrink the insignificant regression coefficients in $\boldsymbol{\beta}_k$ and estimate $\boldsymbol{\theta}$, we introduce penalty term to (1) and optimize the following penalized log likelihood function;

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} \mathcal{L}(\boldsymbol{\theta}) - P_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) \tag{3}$$

where $\mathcal{L}(\boldsymbol{\theta})$ denotes the observed log likelihood, and $P_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) : \mathcal{R}^P \to \mathcal{R}$ is a regularizer of the regression coefficients, and the penalty for each component is dependent on a component specific hyperparameter $\lambda_k > 0$. For notational convenience, we define $\mathcal{L}^p(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}) - P_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$. Various types of penalty were used in mixture regression model [27], [23], but we consider LASSO penalty form as it is convex and thus advantageous for numerical computation [11], i.e.,

$$P_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \Sigma_{k=1}^K \pi_k \Sigma_{j=1}^P \lambda_k |\beta_{jk}| \tag{4}$$

Similar to the case of low dimensional mixture regression, EM algorithm could be adopted by maximizing the penalized complete log likelihood function

$$\mathcal{L}^{pc}(\boldsymbol{\theta}) = \mathcal{L}^c(\boldsymbol{\theta}) - P_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$$

by iterating between the following E-step and M-step:
E-step: computing the conditional expectation of $\mathcal{L}^{pc}(\boldsymbol{\theta})$ with respect to z_{ik} given the current estimates $\boldsymbol{\theta}^{(m)}$. The conditional expectation is

$$Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(m)}) = \Sigma_{i=1}^N \Sigma_{k=1}^K p_{ik}^{(m)} [\log \pi_k + \log \mathcal{N}(y_i; \boldsymbol{x}_i^T \boldsymbol{\beta}_k, \sigma_k^2)] - P_{\boldsymbol{\lambda}}(\boldsymbol{\theta}) \tag{5}$$

Then the conditional expectation of z_{ik} is given by

$$p_{ik}^{(m)} = \frac{\pi_k^{(m)} \mathcal{N}(y_i; {\boldsymbol{\beta}_k^{(m)}}^T \boldsymbol{x}_i, \sigma_k^{2^{(m)}})}{\sum_{l=1}^K \pi_l^{(m)} \mathcal{N}(y_i; {\boldsymbol{\beta}_l^{(m)}}^T \boldsymbol{x}_i, \sigma_l^{2^{(m)}})}$$

M-step: maximizing $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(m)})$ with respect to $\boldsymbol{\theta}$, i.e.,

$$\boldsymbol{\theta}^{(m+1)} = \{\pi_k^{m+1}, \boldsymbol{\beta}_k^{(m+1)}, \sigma_k^{2^{(m+1)}}\}_{k=1}^K = \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(m)}) \tag{6}$$

Unlike the low-dimensional case, where $\pi_k^{m+1}, \boldsymbol{\beta}_k^{(m+1)}, \sigma_k^{2^{(m+1)}}$ all have closed form solutions, maximizing (6) is more complicated due to the involvement of $\pi_k, \boldsymbol{\beta}_k$ in the penalty term and the non-differentiable form of $P_{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ at $\beta_{jk} = 0$.

B. The classification EM algorithm

The Classification Expection Maximization (CEM) algrorithm is an variant of the EM algorithm. It has been popularly used in the Finite Gaussian Mixture Model[28], [29], and shown to be have faster convergence rate [30]. Basically, the assignments $\{z_i\}_{i=1}^N$ define a partition $\mathcal{C} = \bigcup_{k=1}^K C_k$ s.t. $i \in C_k$ iff $z_i = k$. The CEM algorithms maximizes $\mathcal{L}^c(\boldsymbol{\theta})$ through iterating among three steps:

E-Step: calculating conditional expectation of $p_{ik}^{(m)}$, similar to the traditional EM.
C-Step: disentangle the observations into K classes, by assigning $C_k^{(m+1)}$ as the set of observations most likely in cluster k, i.e., $\{i|k = \underset{l\in\{1,...,K\}}{\text{argmax}}\ p_{il}^{(m)}, i = 1, ..., N\}$. Let n_k denotes the total number of observations in cluster k.
M-Step: parameter estimation within each disentangled cluster, where $\pi_k^{(m+1)}$ is estimated as $n_k^{(m+1)}/N$, and $\boldsymbol{\beta}_k^{(m+1)}, \sigma_k^{2^{(m+1)}}$ are simply estimated as the ordinary least square (OLS) estimators using observations in $C_k^{(m+1)}$ only.

We show the convergence of the CEM algorithm for the low-dimensional case in Theorem 1.

Theorem 1. *For the sequence $\mathcal{C}^{(m)}, \boldsymbol{\theta}^{(m)}$ updated as CEM, the complete data likelihood converges to a stationary value. Moreover, if the maximum likelihood estimates of the parameters are well-defined, the sequence $\mathcal{C}^{(m)}, \boldsymbol{\theta}^{(m)}$ also converges to a stationary position.*

The biggest advantage of the CEM algorithm is that it disentangles the mixture into individual non-overlapping components, such that flexible sparsity control could be easily achievable within each component. Hence for the high dimensional mixture regression problem, we could simply replace the OLS estimator in the M step of the CEM algorithm by a sparse estimator, i.e.,

$$\underset{\boldsymbol{\beta}_k, \sigma_k^2}{\textbf{argmax}} \sum_{i\in\mathcal{C}_k^{(m+1)}} \log \mathcal{N}(y_i; \boldsymbol{x}_i^T\boldsymbol{\beta}_k, \sigma_k^2) - \lambda_k\pi_k\Sigma_{j=1}^P|\beta_{jk}|$$

This is simply L_1 regularized linear regression, for which many efficient algorithms exist [31].

C. The CSMR algorithm

Here we proposed the CSMR algorithm to solve the high dimensional mixture regression problem based on the CEM algorithm. In CSMR, the mixture regression setting could handle the hidden cluster problem, and the disentangled clusters under CEM could efficiently solve the feature selection problem in high dimensional setting. At E-step, we calculate the posterior probability p_{ik} similar to traditional EM and ECM; at C-step, we assign each observation to a cluster that it most likely belongs to, similar to traditional CEM; at the M-step, for each component, we perform regularized linear regression to obtain a sparse set of non-zero coefficients.

A big challenge with the penalized mixture regression problem is the choice of component specific penalty parameters λ_k. The λ_k's are related to the amount of regularization, and their selection is a critical issue in a penalized likelihood approach. It is usually based on a trade-off between bias and variance: large values of tuning parameters tend to select a simple model whose parameters estimates have smaller variance, whereas small values of the tuning parameters lead to complex models, with smaller bias. Cross-validation over a grid search is the commonly adopted method to select the optimal combination of λ_k, but this becomes increasingly prohibitive with the increase of K, especially when we don't have a good knowledge of the theoretical range of the λ_k.

Hence, instead of first performing penalized linear regression for given λ_k and then search the optimal combination of λ_k [23], we propose to conduct the tuning of λ_k with cross validation inside the ECM iterations. Specifically, under the CEM algorithm, all the components are disentangled, we could hence perform hyperparameter tuning inside each iteration within each component. This is to say, at the M-step, we not only estimate the regression coefficients, but also find the best tuning parameter λ_k for the component. Hence, at the end of the algorithm, we avoid the hyperparameter tuning, as they have already been selected within the iteration. We adopted efficient cross validation algorithm for selecting the optimal regularization parameter under L_1 regularized linear regression [31]. Since we no longer need to run the algorithm multiple times on a K-dimensional grid space of the penalty parameters, and could hence largely reduce the computational cost. We have shown in simulation studies that penalty parameters selected this way empirically worked very well.

Another adaptation on the traditional ECM algorithm of CSMR is a model refit step following the ECM steps. To increase the numerical stability and achieve faster convergence, at the end of each iteration, we refit the mixture regression model using flexible EM algorithm with only the selected variables of each component. Basically, for each component, the coefficients of the variables not selected at the M-step will be forced to be zero. This could be easily achievable by allowing only the selected variables of component k to enter into the model fitting of the k-th regression parameters.

Algorithm 1 CSMR

```
Input: $X_{N\times P}, Y_{N\times 1}, K$
Output: $\boldsymbol{\theta}, \mathcal{C} = \bigcup_{k=1}^K \mathcal{C}_k, \{\beta_{0k}, \boldsymbol{\beta}_k\}_{k=1}^K$
Initialization: $\boldsymbol{\theta}^{(0)} = \{\pi_k^0, \boldsymbol{\beta}_k^{(0)}, \sigma_k^{2^{(0)}}\}_{k=1}^K$
for m=0,...,Max Iteration do
    E-step: Compute the conditional expectation of $z_{ik}$ similar to traditional EM algorithm.
    C-step: For $k = 1, ..., K$, assign $C_k^{(m+1)}$ as the set of observations that are mostly likely in component $k$.
    M-step: For $k = 1, ..., K$, the relative cluster size is updated by $\hat{\pi}_k^{(m+1)} = \frac{n_k^{(m+1)}}{N}$, and the tuning parameter $\lambda_k^{(m+1)}$, and regression parameters $(\hat{\boldsymbol{\beta}}_k^{(m+1)}, \hat{\sigma}_k^{(m+1)})$ are selected and estimated using cross validation, such as the cv.glmnet function in glmnet package.
    Model refit: refit the FMGR model by allowing only the selected variables in each component and to obtain $\{\pi_k^{(m+1)}, \boldsymbol{\beta}_k^{(m+1)}, \sigma_k^{(m+1)}\}_{k=1}^K$ given by this flexible modeling
    Stop if converged.
end
```

The CSMR algorithm requires the initialized values $\boldsymbol{\theta}$. Here, we order the features based on its individual Pearson correlation with the response variable, and then fit a low-dimensional mixture regression model solved by traditional EM algorithm using the top correlated genes. CSMR is implemented in R, and was made available in https://github.com/zcslab/CSMR.

D. Selection of component number K

The number of clusters K is a sensible parameter because it describes the heterogeneity of the population. For selection of K, we could use a modified BIC criterior that minimizes

$$BIC(K) = -2\mathcal{L}^{pc}(\boldsymbol{\theta}^*_K) + log(N)d_K$$

where $\boldsymbol{\theta}^*_K$ represents the parameter estimates for K, and $d_K = K + (K-1) + \Sigma_{k=1}^K \Sigma_{j=1}^P 1_{\{\beta_{jk}\neq 0\}}$ is the effective number of parameters to be estimated, similar to [32]. Specifically, there are K standard deviations, σ_k, associated with the K regression lines; $K-1$ component proportions, π_k, since $\Sigma_k \pi_k = 1$; and all the non-zero linear regression coefficients for all the K components.

In addition to the BIC criteria, we also offer a cross validation algorithm for the selection of K. Take a 5-fold cross validation as an example. For given K, at each repetition, 80% samples are used for training to obtain the regularized parameter $\boldsymbol{\theta}^*_K$. Then, for a sample $(\boldsymbol{x}_i, y_i)$ drawn from the 20% testing samples, its cluster membership, k_0, is first predicted as

$$k_0 = \max_k \pi^*_{k,K}\mathcal{N}(y_i; \boldsymbol{x}_i^T \boldsymbol{\beta}^*_{k,K}, \sigma^{2*}_{k,K})$$

Here, $\pi^*_{k,K}, \boldsymbol{\beta}^*_{k,K}, \sigma^{2*}_{k,K}$ denote the CSMR estimated parameters when the number of components is K. After assigning the observation to component k_0, we could make prediction of the response based on linear regression, i.e. $\hat{y_i} = \boldsymbol{x}_i^T \boldsymbol{\beta}^*_{k_0,K}$, as well as the associated residual, $y_i - \hat{y_i}$. Notably, such a prediction of the response is different from simple linear regression, as the prediction process requires knowing the value of the response, in order to assign it to the right cluster. After knowing its cluster membership, a prediction of the response could be made.

A large K will tend to overfit the data with more complex model of higher variance, while smaller K might select a simpler model with larger bias. Using the independent testing data, we could decide how to balance the trade-off between bias and variance. To evaluate how the estimated model under K explains the testing data, we could calculate the root-mean-square-error between y_i and $\hat{y_i}$, or Pearson correlation between the two. By repeating this procedure for multiple times, a more robust and stable evaluation of the choice of K should be derived based on the summarized RMSE or Pearson correlations.

IV. Application to simulation data

A. Data generation procedure

We simulated the independent variables $x_i, i = 1, ..., P$, which follows i.i.d normal distribution, i.e., $x_{ij} \sim N(0,1)$. The component proportions were made to be equal, i.e., $\pi_k = \frac{1}{K}$. For component k, a random sample of size M_0 were taken from $\{1, ..., P\}$, denoted as I_k. And $\beta_{ki} \in Unif((-b,-a)\bigcup(a,b))$, if $\beta_{ki} \in I_k$; $\beta_{ki} = 0$, if $\beta_{ki} \notin I_k$.

The response variable Y_k was generated by the following two-step process:
1. Draw component $z_i \in \{1, ..., K\}$ with probability $p(z_i = k|\theta) = \pi_k$.
2. Draw an observation y_i according to normal distribution $N(\beta_{0k} + \boldsymbol{\beta_k^T x_i}, \sigma_k^2)$.

Here, we fix $a = 2, b = 5, P = 100$. We explored the performances of existing methods under 12 different simulation scenarios, for each of which, 100 repetitions were conducted:
Cases 1-3. $N = 200, 300, 400, P = 100, K = 2, \sigma = 1, M_0 = 5$
Cases 4-6. $N = 400, P = 100, K = 2, 3, 4, \sigma = 1, M_0 = 5$
Cases 7-9. $N = 400, P = 100, K = 2, \sigma = 0.5, 1, 2, M_0 = 5$
Cases 10-12. $N = 400, P = 100, K = 2, \sigma = 1, M_0 = 5, 8, 20$

B. Baseline methods

We compared CSMR with five different methods, including $\mathcal{L}_1$ penalized regression, or LASSO; $\mathcal{L}_2$ penalized regression, or Ridge regression (RIDGE); random forest based regression (RF), FMRS [33] and ICC [27]. They differ in their ability to perform prediction, clustering and variable selection, as shown in Table 1.

TABLE I: Baseline methods

	Prediction	Clustering	Variable selection
CSMR	×	×	×
LASSO	×		×
RIDGE	×		
RF	×		×
FMRS	×	×	×
ICC	×	×	×

Among them, CSMR, ICC and FMRS are capable of doing variable selection at the same time of sample clustering. However, FMRS can only deal with relatively lower dimensional features.

C. Performance comparisons

We focused on four metrics for method comparisons: 1) the average correlation between predicted and observed response; 2) the true positive rate (TPR) and 3) true negative rate (TNR) of variable selection; 4) the rand index of sample clustering (RI). Note that for observation i, its predicted response is given by $\Sigma_{k=1}^K z_{ik}(\boldsymbol{x}_i^T \boldsymbol{\beta}_k)$, where z_{ik} is its cluster membership indicator. The average of the four metrics over 100 simulations in each scenario was calculated and shown in Table 2. Here, we assume that the true K is known.

In general, CSMR performs the best in terms of the four evaluation metrics in the majority of the scenarios. For prediction accuracy of the response using correlation, CSMR and

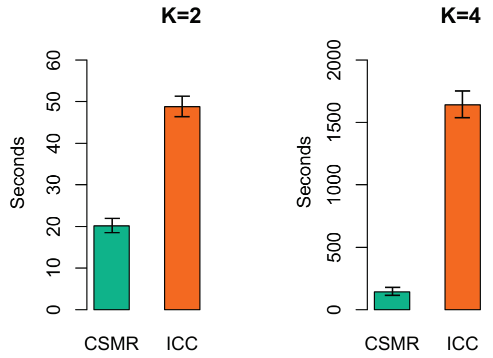

Fig. 2: Time consumption of CSMR, and ICC on simulation data for $K = 2$ (left) and $K = 4$ (right), and $N = 400, \sigma = 1, M_0 = 20$ over 100 repetitions, error bars indicate standard deviations.

ICC perform comparably well, and CSMR slightly better in most of the cases. This is expected as LASSO, RIDGE and RF can not deal with the sample heterogeneity, and FMRS does not work well when the feature dimension is high. For sensitivity and specificity of the variable selection, CSMR performs significantly better than ICC and FMRS. Selection of the right variables is very important as it characterizes the unique features of each component, based on which, we could further deduce the biological interpretation of each unique component. ICC and FMRS suffer from very low sensitivity of variable selection in almost all cases, and their specificity metrics are not desirable either. For clustering, CSMR again has the best or close to the best performance compared with ICC and FMRS. ICC achieved similar performance with CSMR in some cases, but it clearly suffers when K or the number of effective variables M_0 become large. We also compared the computational efficiency of CSMR and ICC under the parameter setting: $N = 400, P = 100, \sigma = 1, M_0 = 20, K$ =2 or 4. Figure 2 shows the computational cost and its standard deviation for two algorithms over 100 repetitions. Clearly, the computational efficiency of ICC drops significantly when K increases from 2 to 4, while the time consumption for CSMR stays approximately the same.

Hence from simulation data, we could see that CSMR achieved the most desirable performance in terms of prediction accuracy, variable selection and clustering, compared with three non-mixture regularized models, and two mixture models. While ICC is competitive in some cases, it severely suffers from poor variable selection, and its computational cost is too prohibitive compared with CSMR. The CSMR has a built-in cross validation step within the CEM iteractions, which could largely increase the sensitivity and specificity of the variable selection procedure, and the flexible model refit step following the CEM steps guarantees that the algorithm could achieve faster convergence and more stable results.

V. Application to CCLE data

A. Description of the dataset

Over the past three decades, the use of molecular data to inform drug discovery and development pipeline has generated huge excitement. Predicting the drug sensitivity becomes an integral part of the precision health initiative. Although earlier efforts successfully identified many new drug targets, the overall clinical efficacy of the developed drugs has remained unimpressive, owing in large part to the population heterogeneity, that is, different patients may have different disease causing factors, and hence drug targets. Here, we apply CSMR to study the patient heterogeneity in their response to different drug treatments, and select the most key molecular features that underlie the heterogeneous disease causes.

We collected gene expression data of 470 cell lines on 7902 genes, as well as the cell lines' sensitivity score to all 24 drugs, from the Cancer Cell Line Encyclopedia (CCLE) dataset [6]. The sensitivity score, or called the AUCC score, is defined as the area above the fitted dose response curve, and it has been shown to have better predictive accuracy of sensitivity to targeted therapeutic strategies than other measures, such as IC50 or EC50 [34]. We applied all five methods on the dataset, where the drug sensitivity score was treated as response variable and the gene expressions as independent variables. Here, FMRS is not applicable as the feature dimension is too high while the sample size is too small, hence it is omitted from further analysis. Our goal is to study the biological mechanism of possible heterogeneity in drug sensitivity, under the hypothesis that cells exhibit subgroup characteristics by selecting different genes that confer their different levels of drug sensitivity.

B. Results

We compare the performances of the five methods using cross validation. Basically, for each drug, we conduct a 5-fold cross validation by holding 80% of the data as training, and 20% as testing data, for each of the 100 repetitions. At each repetition, the 20% testing data is used to independently evaluate the performance of each method. At the training phase, we start by fixing the hyper parameters involved in all methods. The penalty parameters for LASSO and RIDGE were selected by cross validation within the training samples. For RF, the default parameters were used in the function 'randomForest' of the package with the same name. For ICC, we used the selected component number as in its original paper [27]. For CSMR, to select the best K, we performed both cross validation and the traditional BIC criteria introduced in Methods, over a grid of $K = 1, 2, 3, 4, 5, 6$. We adopted the results from cross validation, as there is a lack of rigorous theoretical foundation for the validity of the traditional BIC under this high dimensional setting, and the data driven selection of cross validation seems more reasonable. The selected K for BIC and cross validation using CSMR and used K for ICC is summarized in Supplementary Table S1. With the hyper parameters fixed, we then conduct parameter estimations

TABLE II: Comparisons of CSMR with other five methods in various simulation settings

Metrics	Experiment	$\sigma = 1, N = 400, M_0 = 5$			$K = 2, N = 400, M_0 = 5$			$K = 2, \sigma = 1, M_0 = 5$			$K = 2, \sigma = 1, N = 400$		
		K			σ			N			M_0		
	Parameter	2	3	4	0.5	1	2	200	300	400	5	8	20
$Cor(y, \hat{y})$	**CSMR**	**0.992**	**0.988**	**0.999**	**0.998**	**0.994**	0.977	0.987	**0.993**	**0.994**	**0.992**	**0.995**	**0.994**
	ICC	0.992	0.985	0.909	0.998	0.984	**0.982**	**0.992**	0.992	0.984	0.992	0.994	0.984
	LASSO	0.743	0.654	0.585	0.745	0.778	0.729	0.776	0.756	0.778	0.743	0.754	0.778
	RIDGE	0.784	0.697	0.639	0.783	0.789	0.772	0.834	0.802	0.789	0.784	0.782	0.789
	RF	0.716	0.583	0.487	0.719	0.605	0.700	0.717	0.720	0.605	0.716	0.691	0.605
	FMRS	0.780	0.676	0.568	0.780	0.706	0.769	0.727	0.797	0.706	0.780	0.780	0.706
Variable Selection (TPR)	**CSMR**	**0.999**	**0.950**	**0.538**	**1**	**0.980**	**1**	**0.956**	**1**	**0.980**	**0.999**	**0.998**	**0.980**
	ICC	0.500	0.332	0.339	0.500	0.461	0.500	0.500	0.500	0.461	0.500	0.496	0.461
	FMRS	0.679	0.552	0.487	0.681	0.579	0.674	0.672	0.706	0.579	0.679	0.635	0.500
Variable Selection (TNR)	**CSMR**	**0.993**	**0.976**	**0.785**	**0.994**	**0.992**	**0.968**	**0.966**	**0.990**	**0.992**	**0.993**	**0.992**	**0.992**
	ICC	0.972	0.957	0.669	0.973	0.870	0.735	0.966	0.972	0.870	0.972	0.953	0.870
	FMRS	0.499	0.680	0.758	0.504	0.512	0.500	0.502	0.506	0.512	0.499	0.515	0.500
Sample Clustering (RI)	**CSMR**	**0.893**	0.833	**0.624**	**0.943**	**0.917**	**0.787**	0.852	**0.886**	**0.917**	**0.893**	**0.908**	**0.917**
	ICC	0.887	**0.838**	0.549	0.941	0.879	0.787	**0.878**	0.881	0.879	0.887	0.903	0.879
	FMRS	0.501	0.546	0.624	0.502	0.513	0.502	0.501	0.501	0.513	0.501	0.502	0.513

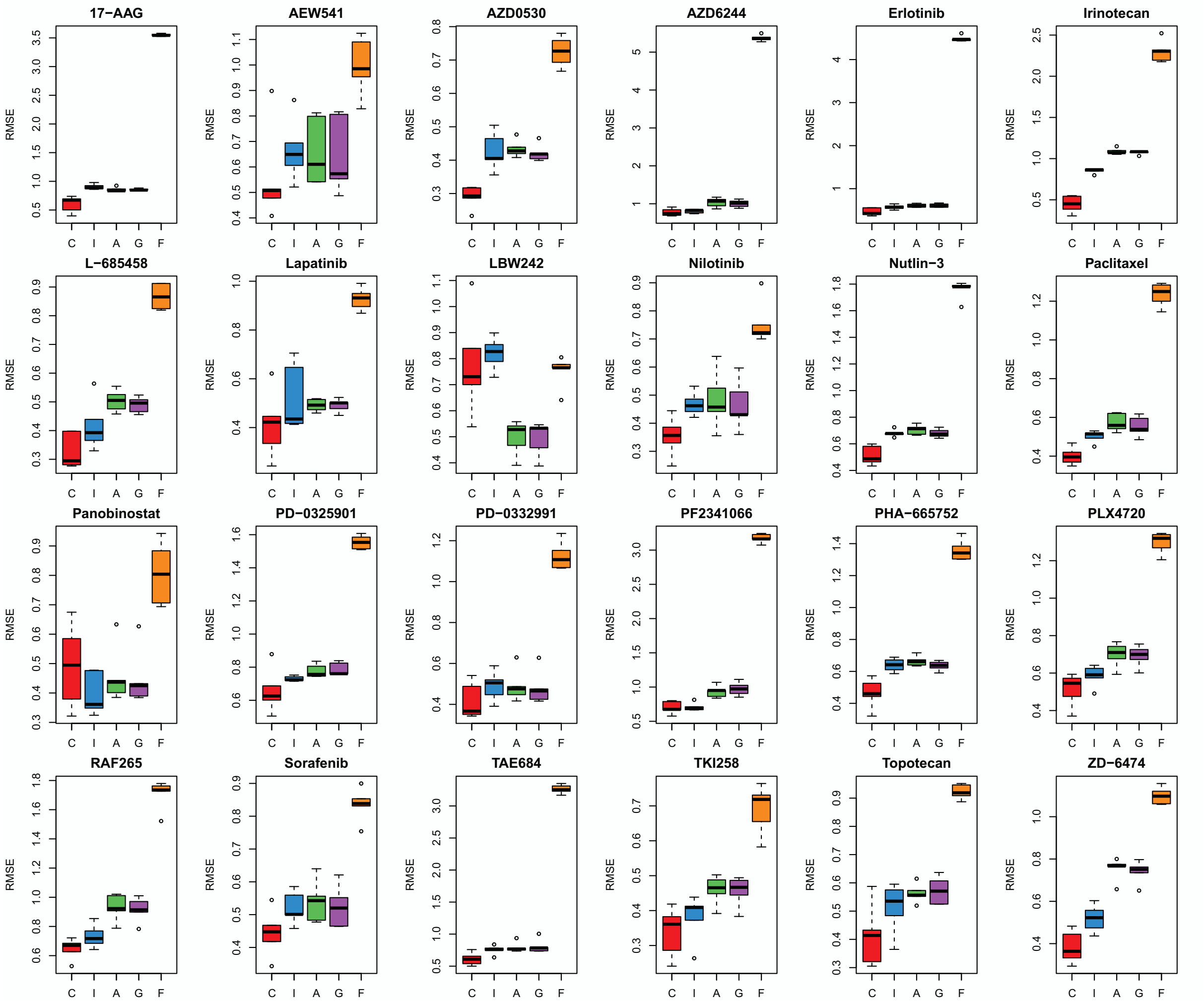

Fig. 3: The distributions of the RMSE over 100 repetitions for the five methods, for the 24 drugs. The lower RMSE value, the better performance. 'C','I','A','G','F' stand for 'CSMR','ICC','LASSO','RIDGE','Random Forest'

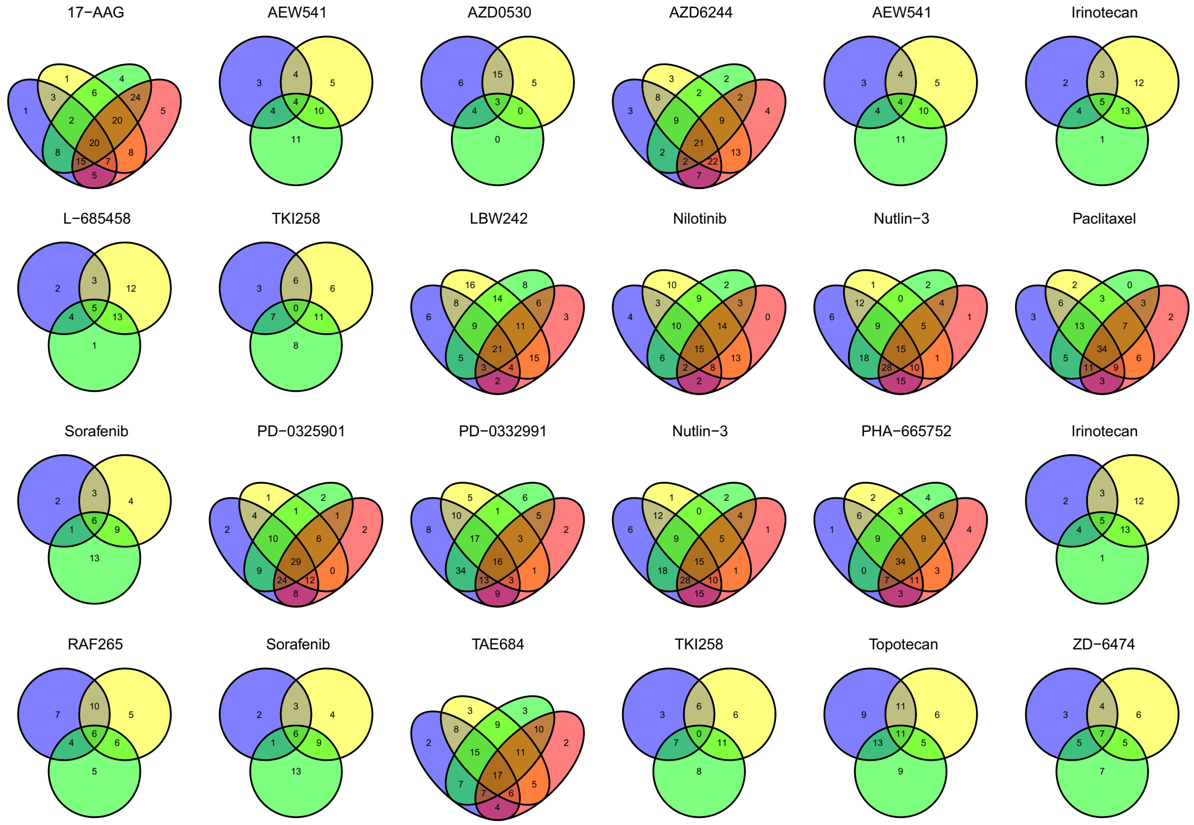

Fig. 4: For each drug, the Venn diagram of the selected genes for different mixing components are shown. The numbers show the size of overlap between the gene sets.

for each of the five methods using the training samples, and concludes the training phase.

At the testing phase, the predicted and true drug sensitivity scores were examined in terms of their correlation, and residual mean squared error (RMSE). Note that this part of the testing data has never been used in the hyper parameter tuning or parameter estimation before. The distributions of RMSE and correlations over 100 repetitions for all the 24 drugs for all the five methods were shown in Figure 3 and Supplementary Figure S1, respectively. For 22 drugs, CSMR had the significantly smaller average RMSE, and was very close to the smallest RMSE for the rest of the two drugs; and we could make the same conclusions based on the correlation results as well. This demonstrated the consistent and robust performance of CSMR over the others.

Among the five methods, RF had the poorest performance on the testing data, probably caused by model overfitting. LASSO and RIDGE worked much better than RF, probably due to its power in model selection. However, they performed significantly worse than ICC and CSMR in majority of the cases, which indicates the existence of population heterogeneity and necessity of using mixture modeling. The performance of ICC is much worse than CSMR in most of the cases, which we believe is caused by the under-estimation of the population heterogeneity by ICC. In other words, the selection of K in ICC is too conservative. In fact, according to cross validation, the number of distinct clusters given by CSMR for the drugs is either 3 or 4, while for ICC, the number of distinct clusters are determined to be less than 3 for half of the drugs. We believe that cross validation is a data driven approach for selection of K, and should be more reasonable than theoretically derived criteria. In the case of CCLE data, the samples are different types of cells from very different experimental and genetic backgrounds, and it is expected that they would pick up different molecular mechanisms to deal with the attacks of the drugs. Hence, the cluster number given by CSMR is more realistic than ICC. It is wrothy of note that for those drugs that CSMR and ICC gave the same number of distinct clusters, namely Irinotecan, L-685458, Lapatinib, Paclitaxel, PD-0332991, PHA-665752 and TKI258, CSMR exhibited much smaller RMSE than ICC.

Figure 4 demonstrated the Venn diagram of the selected genes for different components for each drug, and all the selected genes could be found in Supplementary Table S2. It could be seen that for the same drug, different clusters of cells indeed acquire different coping mechanisms, as seen by the different set of genes selected. This again confirms the high heterogeneous populations within the CCLE cohort. For each drug, we pooled all the selected genes together and conducted pathway enrichment analysis against 1,328 pathways collected in [35], and the top enriched pathways are shown in Supplementary Figure S2. Again, it could be seen that different responses to different drugs have been employed.

VI. Conclusions

With the recent rapid evolution in genomic technologies, we have now entered a new phase, one in which it is possible to comprehensively characterize the molecular profiles of large population of subjects. Importantly, the development of sequencing technologies has been paired with a transition towards integrating molecular data with phenotypic data, such as in the electronic medical records. Such a synergy has the potential to ultimately facilitate the generation of a data commons useful for identifying relationships between molecular variations and their clinical presentations. Unfortunately, existing big data analysis tools for mining the information rich data commons has not been very impressive with regards to the overall transnational or clinical efficacy, owing in large part to the heterogeneous causes of disease. It is hence imperative to unveil the relationship between the molecular manifestations and the clinical presentations, while taking into account the possible heterogeneity of the study subjects.

In this paper, we proposed a novel supervised clustering algorithm using penalized mixture regression model, called CSMR, to deal with the challenges in studying the heterogeneous relationships between high dimensional molecular features to a phenotype. CSMR is capable of simultaneous stratification of the sample population and sparse feature-wise characterization of the subgroups. The algorithm was adapted from the classification expectation maximization algorithm, which offers a novel supervised solution to the clustering problem, with substantial improvement on both the computational efficiency and biological interpretability. Experimental evaluation on simulated benchmark datasets with different settings demonstrated that the CSMR can accurately identify the subspaces on which subset of features are explanatory to the response variables and the feature characteristics of the subspaces, and it outperformed the baseline methods. Application of CSMR on the heterogeneous CCLE dataset demonstrated the superior performance of CSMR over the others. On the CCLE dataset, CSMR is powerful in recapitulating the distinct subgroups hidden in the pool of cell lines with regards to their coping mechanisms to different drugs. CSMR also demonstrated the uniqueness of different subgroups for the same drug, as seen by the distinctly selected genes for the subgroups.

In summary, CSMR represents a big data analysis tool with the potential to bridge the gap between advancements in biotechnology and our understanding of the disease, and resolve the complexity of translating the clinical manifestations of the disease to the real causes underpinning it. We believe that such a tool will bring new understanding to the molecular basis of a disease, and could be of special relevance in the growing field of personalized medicine.

References

[1] Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.

[2] Andreas Schlicker, Garry Beran, Christine M Chresta, Gael McWalter, Alison Pritchard, Susie Weston, Sarah Runswick, Sara Davenport, Kerry Heathcote, Denis Alferez Castro, et al. Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. *BMC medical genomics*, 5(1):66, 2012.

[3] Justin Guinney, Rodrigo Dienstmann, Xin Wang, Aurélien De Reyniès, Andreas Schlicker, Charlotte Soneson, Laetitia Marisa, Paul Roepman, Gift Nyamundanda, Paolo Angelino, et al. The consensus molecular subtypes of colorectal cancer. *Nature medicine*, 21(11):1350–1356, 2015.

[4] Andriy Marusyk and Kornelia Polyak. Tumor heterogeneity: causes and consequences. *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer*, 1805(1):105–117, 2010.

[5] Martin Köbel, Steve E Kalloger, Niki Boyd, Steven McKinney, Erika Mehl, Chana Palmer, Samuel Leung, Nathan J Bowen, Diana N Ionescu, Ashish Rajput, et al. Ovarian carcinoma subtypes are different diseases: implications for biomarker studies. *PLoS medicine*, 5(12):e232, 2008.

[6] Jordi Barretina, Giordano Caponigro, Nicolas Stransky, Kavitha Venkatesan, Adam A Margolin, Sungjoon Kim, Christopher J Wilson, Joseph Lehár, Gregory V Kryukov, Dmitriy Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012.

[7] Amrita Basu, Nicole E Bodycombe, Jaime H Cheah, Edmund V Price, Ke Liu, Giannina I Schaefer, Richard Y Ebright, Michelle L Stewart, Daisuke Ito, Stephanie Wang, et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell*, 154(5):1151–1161, 2013.

[8] Wanjuan Yang, Jorge Soares, Patricia Greninger, Elena J Edelman, Howard Lightfoot, Simon Forbes, Nidhi Bindal, Dave Beare, James A Smith, I Richard Thompson, et al. Genomics of drug sensitivity in cancer (gdsc): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic acids research*, 41(D1):D955–D961, 2012.

[9] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, Joshua M Stuart, Cancer Genome Atlas Research Network, et al. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113, 2013.

[10] Jianqing Fan, Fang Han, and Han Liu. Challenges of big data analysis. *National science review*, 1(2):293–314, 2014.

[11] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[12] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

[13] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[14] Geoffrey J McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

[15] Chris Fraley and Adrian E Raftery. How many clusters? which clustering method? answers via model-based cluster analysis. *The computer journal*, 41(8):578–588, 1998.

[16] Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631, 2002.

[17] Stephen Goldfeld and Richard Quandt. The estimation of structural shifts by switching regressions. In *Annals of Economic and Social Measurement, Volume 2, number 4*, pages 475–485. NBER, 1973.

[18] Dankmar Böhning. *Computer-assisted analysis of mixtures and applications: meta-analysis, disease mapping and others*, volume 81. CRC press, 1999.

[19] Christian Hennig. Identifiablity of models for clusterwise linear regression. *Journal of Classification*, 17(2), 2000.

[20] Wenxin Jiang and Martin A Tanner. Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation. *Annals of Statistics*, pages 987–1011, 1999.

[21] Lei Xu and Michael I Jordan. On convergence properties of the em algorithm for gaussian mixtures. *Neural computation*, 8(1):129–151, 1996.

[22] Sylvia Frühwirth-Schnatter. *Finite mixture and Markov switching models*. Springer Science & Business Media, 2006.

[23] Abbas Khalili and Jiahua Chen. Variable selection in finite mixture of regression models. *Journal of the american Statistical association*, 102(479):1025–1038, 2007.

[24] Nicolas Städler, Peter Bühlmann, and Sara Van De Geer. L1-penalization for mixture regression models. *Test*, 19(2):209–256, 2010.

[25] Jianqing Fan and Jinchi Lv. Comments on: L1-penalization for mixture regression models. *Test*, 19(2):264–269, 2010.

[26] Luke R Lloyd-Jones, Hien D Nguyen, and Geoffrey J McLachlan. A globally convergent algorithm for lasso-penalized mixture of linear regression models. *Computational Statistics & Data Analysis*, 119:19–38, 2018.

[27] Qianyun Li, Runmin Shi, and Faming Liang. Drug sensitivity prediction with high-dimensional mixture regression. *PloS one*, 14(2), 2019.

[28] Johannes Blömer, Sascha Brauer, and Kathrin Bujna. Hard-clustering with gaussian mixture models. *arXiv preprint arXiv:1603.06478*, 2016.

[29] Gilles Celeux and Gérard Govaert. A classification em algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis*, 14(3):315–332, 1992.

[30] Susana Faria and Gilda Soromenho. Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation*, 80(2):201–225, 2010.

[31] Trevor Hastie and Junyang Qian. Glmnet vignette. *Retrieved June*, 9(2016):1–30, 2014.

[32] Wei Pan and Xiaotong Shen. Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, 8(May):1145–1164, 2007.

[33] Abbas Khalili, Jiahua Chen, and Shili Lin. Feature selection in finite mixture of sparse normal linear models in high-dimensional feature space. *Biostatistics*, 12(1):156–172, 2011.

[34] In Sock Jang, Elias Chaibub Neto, Justin Guinney, Stephen H Friend, and Adam A Margolin. Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. In *Biocomputing 2014*, pages 63–74. World Scientific, 2014.

[35] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.