Sparse Group Selection on Fused Lasso Components for Identifying Group-specific DNA Copy Number Variations

Ze Tian*, Huanan Zhang* and Rui Kuang*
*Department of Computer Science and Engineering
University of Minnesota Twin Cities, Minneapolis, Minnesota 55455–0213
Email: tianze@cs.umn.edu, huanan@cs.umn.edu, kuang@cs.umn.edu

***Abstract*—Detecting DNA copy number variations (CNVs) from arrayCGH or genotyping-array data to correlate with cancer outcomes is crucial for understanding the molecular mechanisms underlying cancer. Previous methods either focus on detecting CNVs in each individual patient sample or common CNVs across all the patient samples. These methods ignore the discrepancies introduced by the heterogeneity in the patient samples, which implies that common CNVs might only be shared within some groups of samples instead of all samples. In this paper, we propose a latent feature model that couples sparse sample group selection with fused lasso on CNV components to identify group-specific CNVs. Assuming a given group structure on patient samples by clinical information, sparse group selection on fused lasso (SGS-FL) identifies the optimal latent CNV components, each of which is specific to the samples in one or several groups. The group selection for each CNV component is determined dynamically by an adaptive algorithm to achieve a desired sparsity. Simulation results show that SGS-FL can more accurately identify the latent CNV components when there is a reliable underlying group structure in the samples. In the experiments on arrayCGH breast cancer and bladder cancer datasets, SGS-FL detected CNV regions that are more relevant to cancer, and provided latent feature weights that can be used for better sample classification.**

***Keywords*-sparse group learning; fused lasso; group lasso; DNA copy number variations;**

I. Introduction

There are normally two copies of each gene in the double-stranded DNAs of human genome. Alterations of the DNAs can lead to a different number of copies of the genes in the DNA. If a DNA region is deleted in one or both strands, there will be a fewer number of copies of the genes and on the contrary if a DNA region is duplicated, there will be a larger number of copies of the genes. These events of amplification or deletion of a large DNA segment on chromosomes are called DNA copy number variations (CNVs). It has been confirmed in many recent studies that chromosomal aberrations of DNA copy numbers, rearrangement and structures have association with cancer and other diseases [1]. Among the aberrations, DNA copy number variations (CNVs) are believed to play an important role in tumorigenesis [2].

DNA CNVs can be measured by comparative genomic hybridization (CGH), which compares the copy number of a differentially labeled case sample with a normal reference DNA. ArrayCGH technology based on DNA microarray can currently allow genome-wide identification of CNV regions by CGH measuring at a number of sampled locations at different resolutions [3]. ArrayCGH data provide important information of candidate cancer loci for the classification of patients and discovery of molecular mechanisms of cancers [4]. Thus, one of the main tasks of arrayCGH data analysis is to identify the CNVs represented as amplified and deleted regions on the chromosomes and correlate them with diseases. These regions are expected to encode important structural genomic variations related to the diseases.

II. Related Work

Since CNV data are series of $\log$ intensity ratios at the sampled locations (probes), the adjacent probe locations are more likely to be associated in the same CNV event. For single-sample CNV detection, fused lasso appears to be a promising model [5]. In the fused lasso, L_1-norm is used in the penalty term to smooth the data by encouraging sparsity of the data and also the sparsity of the change points. Specifically, the method finds the segmented series β by solving the optimization problem

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_i (y_i - \beta_i)^2 \right\}$$

subject to $\sum_j |\beta_j| \leq s_1$ and $\sum_j |\beta_j - \beta_{j+1}| \leq s_2$, where y_i is the log2 ratio measurement of the ith probe and β_i is the corresponding value after smoothing. Here, s_1 controls the overall sparsity of CNV regions (the number of nonzero values in β) and s_2 controls the number of CNV alterations (the number of change points between the adjacent probes). By examining the non-zero values in β, CNVs can be identified for the sample. The procedure is repeated for all the samples and the resulted segmented series are aggregated to report the identified common CNVs.

For multi-sample CNV detection, all samples are analyzed simultaneously in one optimization framework to identify the amplification or deletion regions shared across all samples. For example, for p copy-number profiles of length n,

[6] and [7] proposed the following optimization problem

$$\min_{U\in\mathbb{R}^{n\times p}} \|Y-U\|^2 + \lambda \sum_{i=1}^{n-1} \|U_{i+1,\bullet} - U_{i,\bullet}\|,$$

where Y is the $n\times p$ CNV profile matrix, U is the de-noised segmentation approximating Y and $U_{i,\bullet}$ is the ith row of U. A fast group least-angle regression (LARS) algorithm can be applied to solve the optimization framework approximately to detect shared change-points from the multiple CNV profiles. Since the change-points are detected from all profiles in the framework, it is expected to be more accurate than detecting change-points independently from each CNV profile.

Under the same motivation that CNVs are usually shared by multiple samples, instead of approximating the profile matrix Y by a segmentation matrix of the same size, another more advanced modeling is to detect the shared CNVs as latent fused features by low-rank matrix factorization decomposed from Y. For example, the widely used dimensionality reduction method principal component analysis (PCA) can decompose Y into orthogonal principle components. The projection of Y to a low-dimensional space obtains coefficients of the principle components to preserve the variance. However, practically it is not feasible to interpret the principle components as CNVs since the principle components cannot be explained as CNV patterns without fusing the adjacent features with lasso.

More recently, a Fused Lasso Latent Feature Model (FLLat) was proposed by [8] for detecting latent CNV components. For the profile matrix Y with S samples and L probes, FLLat decomposes it as a weighted sum of a fixed number of latent feature components, which are smoothed by fused lasso. The corresponding optimization problem for FLLat is

$$\min_{\Gamma,\Theta} \sum_{s=1}^{S}\sum_{l=1}^{L}\left(Y_{ls} - \sum_{j=1}^{J}\Gamma_{lj}\Theta_{js}\right)^2 + \lambda_1 \sum_{j=1}^{J}\sum_{l=1}^{L}|\Gamma_{lj}| + \lambda_2 \sum_{j=1}^{J}\sum_{l=2}^{L}|\Gamma_{lj} - \Gamma_{l-1,j}|$$

subject to $\sum_{s=1}^{S}\Theta_{js}^2 \le 1$ for each j, where J is the number of latent features and Y_{ls} is the log intensity ratio of the lth probe for the sth sample. This model minimizes the sum of the square errors as well as the fused lasso penalties on the latent feature Γ. It is clear that the model does not assume any structure on the weights of the latent fused lasso components Θ.

III. Method

A. Motivation

It is well acknowledged that there is often group-structured prior information on the samples in cancer genomic datasets, which accounts for the heterogeneities among the patients. For example, the samples can be grouped by different tumor grades or stages, or by survival and metastatic status. The samples in each or some of the groups might be associated with CNV components that are only associated with the samples in the group(s). It is actually known that samples with different phenotypes also show different frequencies of CNVs. For example, low and medium grade tumors of bladder cancer generally contain few changes [9]. Thus, it is more biologically interesting to identify CNV patterns for the samples under the group-structure given by prior information. To achieve this objective, we propose *sparse group selection on fused lasso components* (SGS-FL) for integrating group information on the fused lasso components. SGS-FL assumes a group structure on the component coefficients and attempts to select only a small number of groups for each component. SGS-FL also requires that the coefficients of latent features to be non-negative for better distinguishing CNVs as regions with amplifications or deletions. The outline of SGS-FL is given in Fig. 1.

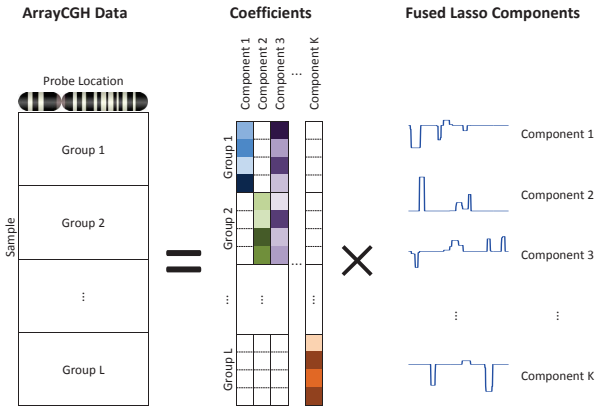


Figure 1. Outline of SGS-FL. The arrayCGH data is factorized into the coefficient matrix and K CNV components. After the factorization, each CNV sample can be reconstructed by the sum of the components weighted by the coefficients. The fused lasso penalty on each component encourage the step-function pattern to model real CNV events. The patient samples are divided into L groups. For each component, only the samples in the selected groups will have nonzero weights. For example, only samples in group 1 have nonzero weights for component 1 and only samples in group 1 and 2 have nonzero weights for component 3. The group selection enforces the sparseness of the coefficients by groups.

B. Notation

We denote the arrayCGH profiles on a chromosome as a $m\times n$ matrix Y where m is the number of samples and n is the number of probes. Y_{ij} is the $\log_2$ intensity ratio measured for the ith sample on the jth probe. Assume that the probes are ordered by their positions on the chromosome. The objective is to decompose Y into a $m\times K$ matrix X and a $K\times n$ matrix W such that $\frac{1}{2}\|Y-XW\|_F^2$ is minimized where K is the number of the components of

latent features. Here W are the components of the latent features and X are the coefficients of the components for all samples. Each sample $Y_{i,\bullet}$ is reconstructed by the linear weighted sum of latent features $\sum_k X_{ik} \cdot W_{k,\bullet}$. We further assume that the m samples are categorized into L disjoint groups as $G = \{g_1, g_2, \ldots, g_L\}$ where $g_l \subset \{1, 2, \ldots, m\}$ is the set of indexes of all samples in the lth group.

C. Regularization Framework

We propose the following optimization problem to minimize reconstruction error and fused lasso under the group selection constraints:

$$\min_{X,W} \frac{1}{2} \|Y - XW\|_F^2 + \lambda_1 \sum_{k=1}^{K} |W_{k,\bullet}| + \lambda_2 \sum_{k=1}^{K} \sum_{j=2}^{n} |W_{kj} - W_{k,j-1}| \quad (1)$$

subject to

$$b_{lk} X_{g_l,k} = 0 \text{ for } l = 1, \ldots, L \text{ and } k = 1, \ldots, K$$
$$X_{ik} \geq 0 \text{ for } i = 1, \ldots, m \text{ and } k = 1, \ldots, K$$
$$\sum_{i=1}^{m} X_{ik}^2 = 1 \text{ for } k = 1, \ldots, K$$

where b_{lk} is a binary indicator variable of selecting the lth group for the kth latent feature component and $X_{g_l,k}$ is a sub-vector of $X_{\bullet,k}$ with indexes $i \in g_l$. If the m samples are divided into L strictly non-overlapping groups, X can be rearranged into a matrix of $L \times K$ vectors as

$$X' = \begin{pmatrix} X_{g_1,1} & \cdots & X_{g_1,K} \\ \vdots & \ddots & \vdots \\ X_{g_L,1} & \cdots & X_{g_L,K} \end{pmatrix}.$$

b_{lk} is introduced for group selection on each fused lasso component $W_{k,\bullet}$. Specifically, if $b_{lk} = 1$, all the coefficients of $X_{g_l,k}$, the kth latent feature component from the lth group, need to be 0; otherwise if $b_{lk} = 0$, the coefficients of $X_{g_l,k}$ can be any nonnegative weights. In other words, group l is selected for component k if $b_{lk} = 0$. The b_{lk} acts as a gating of the weights $X_{g_l,k}$ on the component $W_{k,\bullet}$ and $W_{k,\bullet}$ is only specific to the samples in the selected groups (with $b_{lk} = 0$). In the optimization problem, b_{lk}s are chosen dynamically in each iteration of the optimization algorithm according to a fixed global parameter $r \in [0, 1]$.

D. Sparse Group Selection

Given the parameter $r \in [0, 1]$ and the current weights and components, b_{lk}s are determined for each latent feature component separately. We first define variance factors $v^{(k)} = (Y - X_{\bullet,\neq k} W_{\neq k,\bullet}) W'_{k,\bullet}$ where $X_{\bullet,\neq k}$ is the matrix after removing the kth column from X and $W_{\neq k,\bullet}$ is the matrix after removing the kth row from W. Each $v^{(k)}$ evaluates the importance of the component $W_{k,\bullet}$ to the reconstruction error. The larger the $v^{(k)}$, the more important the $W_{k,\bullet}$. Then, the role of a group l to the component k is evaluated as

$$\gamma_l^{(k)} = \frac{\left\| v_{g_l}^{(k)} \right\|}{\sqrt{|g_l|} \left\| W_{k,\bullet} \right\|^2}, \quad (2)$$

where $|g_l|$ is the cardinality of g_l and $v_{g_l}^{(k)}$ is the sub-vector of $v^{(k)}$ with indexes in g_l. The importance vector for group l is normalized by the size of group l and the 2-norm of component $W_{k,\bullet}$. Then, we can sort the L groups by $\gamma_l^{(k)}$ such that $\gamma_{q_1}^{(k)} \geq \gamma_{q_2}^{(k)} \geq \cdots \geq \gamma_{q_L}^{(k)}$. Based on the ranking of the groups, b_{lk}s are calculated by

$$b_{q_l,k} = \begin{cases} 0 & \text{if } l = 1 \text{ or } \sum_{s=1}^{l-1} \gamma_{q_s}^{(k)} / \sum_{s=1}^{L} \gamma_{q_s}^{(k)} < r \\ 1 & \text{otherwise.} \end{cases} \quad (3)$$

For each component $W_{k,\bullet}$, at least one group is selected and additional groups are selected based on their importance proportional to the total importance. More intuition of group selection by the ranking of $\gamma_{q_l}^{(k)}$ and its connection to group lasso are discussed in section III-F.

E. Alternative Optimization

Eqn. (1) can be solved with alternative optimization to get an empirical solution. We alternate between fixing W and solving for X and vice versa until both W and X do not change anymore in the iterations. The complete SGS-FL algorithm is described in Fig. 2. W is initialized with the first K principle components of Y computed by PCA (line 1). Then, we solve X column by column given W (line 3-14) and solve W row by row given X (line 15-19). The algorithm iterates until both X and W converge. Specifically, for the kth column of X, we first compute $\{b_{lk}\}$ as in Eqn. (3) (line 5-11), and then update $X_{\bullet,k}$ by solving the following sub-optimization problem on line 12:

$$\min_{X_{\bullet,k}} \frac{1}{2} \|Y - XW\|_F^2 \quad (4)$$

subject to $b_{lk} X_{g_l,k} = 0$, $X_{ik} \geq 0$ and $\sum_{i=1}^{m} X_{ik}^2 = 1$. Eqn. (4) can be solved with standard quadratic optimization techniques. This procedure is repeated iteratively for each column of X until X does not change anymore. Similarly, for each row of W, we solve the sub-optimization problem to update $W_{k,\bullet}$,

$$\min_{W_{k,\bullet}} \frac{1}{2} \|Y - XW\|_F^2 + \lambda_1 |W_{k,\bullet}| + \lambda_2 \sum_{j=2}^{n} |W_{kj} - W_{k,j-1}|, \quad (5)$$

which is the fused lasso problem. We used the package provided by [10] in our implementation to solve Eqn. (5) on line 17.

Input: arrayCGH data $Y \in \mathbb{R}^{m\times n}$, the number of latent features $K \in \mathbb{Z}^+$, the group-sparsity-controlling parameter $r \in [0,1]$, the parameters $\lambda_1, \lambda_2 \in \mathbb{R}^+$ for lasso and fused lasso penalties.

Output: The non-negative coefficient matrix $X \in \mathbb{R}^{m\times K}$, the latent feature matrix $W \in \mathbb{R}^{K\times n}$.

Initialize W as the first K principle components of Y and X as 0.
repeat
  repeat
    for $k = 1, \ldots, K$ **do**
      $v^{(k)} \leftarrow (Y - X_{\bullet,\neq k} W_{\neq k,\bullet}) W'_{k,\bullet}$
      for $l = 1, \ldots, L$ **do**
        $\gamma_l^{(k)} \leftarrow \frac{\left\|v_{g_l}^{(k)}\right\|}{\sqrt{|g_l|}\|W_{k,\bullet}\|^2}$
      end for
      for $l = 1, \ldots, L$ **do**
        $b_{lk} = \begin{cases} 0 & \text{if } l = \operatorname{argmax}_s \gamma_s^{(k)} \text{ or} \\ & \sum_{s\in\{l'|\gamma_{l'}^{(k)} > \gamma_l^{(k)}\}} \gamma_s^{(k)} / \sum_{s=1}^{L} \gamma_s^{(k)} < r \\ 1 & \text{otherwise.} \end{cases}$
      end for
      $X_{\bullet,k} \leftarrow \operatorname{argmin}_{X_{\bullet,k}} \frac{1}{2}\|Y - XW\|_F^2$ subject to $b_{lk} X_{g_l,k} = 0$, $X_{ik} \geq 0$ and $\sum_{i=1}^{m} X_{ik}^2 = 1$
    end for
  until X does not change
  repeat
    for $k = 1, \ldots, K$ **do**
      $W_{k,\bullet} \leftarrow \operatorname{argmin}_{W_{k,\bullet}} \frac{1}{2}\|Y - XW\|_F^2 + \lambda_1 |W_{k,\bullet}| + \lambda_2 \sum_{j=2}^{n} |W_{kj} - W_{k,j-1}|$
    end for
  until W does not change
until X and W do not change

Figure 2. SGS-FL algorithm

F. Relation to Group Lasso

[11] proposed a group sparsity regularization method to introduce the group-structured prior knowledge for nonnegative matrix factorization. The objective function of their approach is

$$\min_{W\geq 0, H\geq 0} \frac{1}{2}\|Y - WH\|_F^2 + \alpha\|W\|_F^2 + \beta \sum_{b=1}^{B} \left\|H^{(b)}\right\|_{1,q} \tag{6}$$

where Y is the original data matrix and the coefficient matrix H is divided into B groups as $\{H^{(1)}, \cdots, H^{(B)}\}$ by prior knowledge. The motivation of the regularization term on $\{H^{(b)}\}$ is that samples in the same group are expected to share the same sparsity patterns in their latent factor representation. Eqn. (6) uses a global parameter β on group lasso to enforce the group sparsity. However, it does not fit in the problem of CNV detection since the magnitude of latent features (log ratio intensities) can be in very different scales, and it is not possible to choose a global parameter suitable for all the latent feature components. Moreover, Eqn. (6) does not include the fused lasso penalty, which is necessary for the CNV problem.

Now we examine the following group lasso problem similar to Eqn. (6),

$$\min_{X_{\bullet,k}} \frac{1}{2}\|Y - XW\|_F^2 + \gamma \sum_{l=1}^{L} p_l \|X_{g_l,k}\|_2 ,$$

where $p_l = \sqrt{|g_l|}$ is the weight of the lth group. Note that only $X_{\bullet,k}$ are variables in this problem and all other columns of X are fixed. For this non-overlapping group lasso problem, there exists a $\gamma_l^{(k)}$ for each group g_l such that when $\gamma \geq \gamma_l^{(k)}$, the optimal solution for $X_{g_l,k}$ is zero; when $\gamma < \gamma_l^{(k)}$, the optimal solution for $X_{g_l,k}$ is nonzero [12] and actually,

$$\gamma_l^{(k)} = \frac{\left\|v_{g_l}^{(k)}\right\|}{\sqrt{|g_l|}\|W_{k,\bullet}\|^2}$$

where $v_{g_l}^{(k)}$ is defined the same as in Eqn. (2). Thus, it is exactly the $\gamma_l^{(k)}$ that we used to compute $\{b_{lk}\}$ in Eqn.(3).

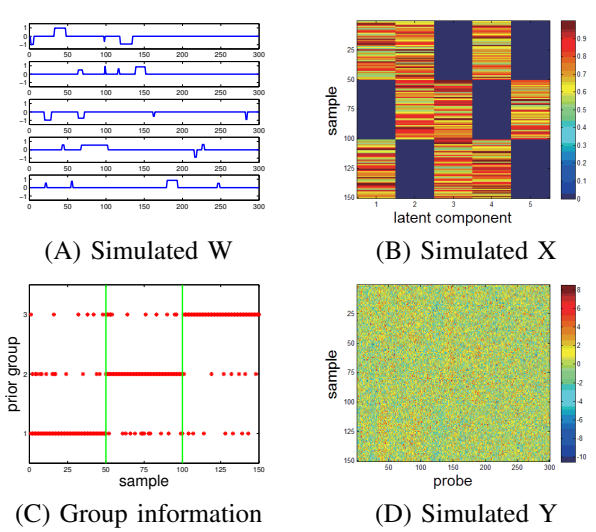
In summary, instead of using group lasso to get sparse X directly, SGS-FL first applies the group lasso setting with the parameter r to adaptively determine $\{b_{lk}\}$. Then, b_{lk}s are used to compute whether $X_{g_l,k}$ should be zero or nonzero in the optimization. Thus, r controls the sparsity of X through $\{b_{lk}\}$. Empirically, using r for adaptive sparse group selection instead of solving a group lasso problem directly for a fixed global parameter γ is more reliable and stable for CNV data analysis.

IV. Simulation

In the simulation, we compare SGS-FL with FLLat on the performance of learning the latent components and the coefficients from an artificial CNV dataset. We also tested the effect of the group sparsity parameter r and evaluated the scalability and the convergency characteristics of SGS-FL.

A. Data Generation

We first generated simulated latent CNV components W and coefficients X with a sparse group structure, and then constructed a CNV dataset $Y = XW + \Xi$, where Ξ are IID gaussian noises, as illustrated in Fig. 3. The simulated arrayCGH dataset contains 150 samples with 300 probes. There are 5 latent components, each of which contains 4 independent events of copy number gain or loss, shown in Fig. 3(A). The 150 samples are equally divided into 3 groups with 50 samples in each group and the corresponding relation between the 3 groups and the 5 components is shown in Fig. 3(B). The group prior is shown in Fig. 3(C). Errors are introduced in the prior information as a certain percentage of misplaced samples in each group. As shown



(A) Simulated W (B) Simulated X

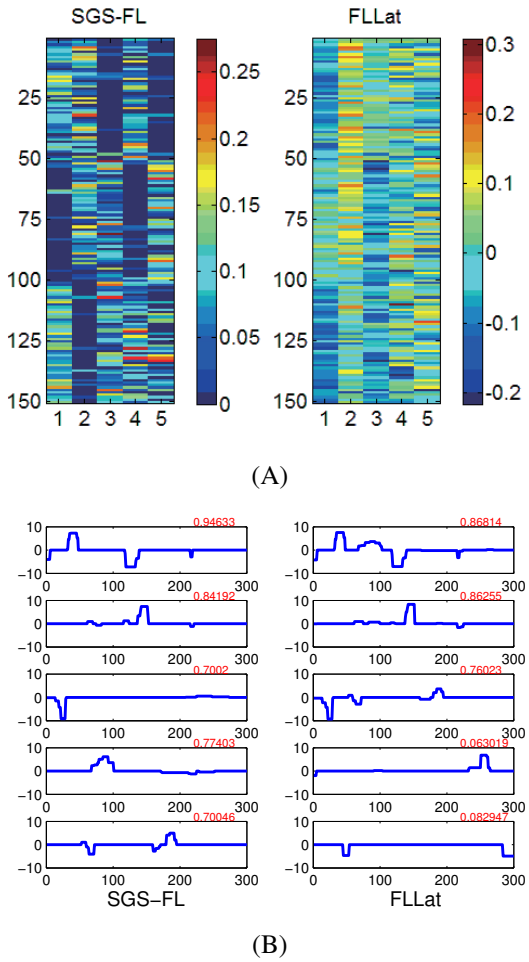
(C) Group information (D) Simulated Y

Figure 3. Simulation data. (A) Each latent CNV component contains four randomly generated copy number gains/losses. (B) The samples are divided into 3 groups of equal size. The coefficients are nonzero only between a group and its corresponding components. (C) Errors are introduced into the prior groups, i.e. a certain percentage of samples are misplaced into the wrong group. (D) The noisy simulated CNV dataset. There is no observable pattern although the data is constructed from the sample groups and the latent CNV components.

in Fig. 3(C), each prior group contains samples from all the three true groups although majority of the samples are from only one group. Given the W and X, the dataset $Y = XW + \Xi$ is shown in Fig. 3(D). Y is a very noisy dataset. k-means clustering with $k = 3$ on the dataset results in error rate above 50%. The objective is to recover W and X from the noisy data Y with SGS-FL and FLLat.

B. Performance of Recovering W and X

Bayesian Information Criterion (BIC) [8] is used to determine the hyper-parameters λ_1 and λ_2 for applying SGS-FL and FLLat. The group sparsity parameter r is set to 0.5. Fig. 4(A) shows the learned hidden components X. Clearly, the X learned by SGS-FL preserves the group structure and is more similar to the original X, compared with the X learned by FLLat. Guided by the prior group information, for each component the coefficients are learned only for the samples in the selected groups. The coefficients in the unselected groups are all zero. For example, for first latent component, only group 1 and 3 are chosen. Fig. 4(B) shows the learned latent components. SGS-FL successfully recovered the 5 latent components with a lowest correlation 0.70 with the original component, while FLLat detected two wrong latent components that are completely different. In the FLLat method, the mistakes in the components can be matched with the wrong weights in X: column 4 and column 5 in the X in Fig. 4(A) do not capture any group relations and thus, the corresponding components are derived to support the wrong samples. On the contrary, the X learned



(A)

(B)

Figure 4. Performance of recovering the latent components and the weight coefficients with SGS-FL and FLLat. (A) The learned coefficient matrices. (B) The learned latent components. The red numbers above each component is the Pearson correlation coefficient between the component and its corresponding original latent component.

by SGS-FL preserves the group structures and the correct samples are used to derive the components. Note that since we introduced noise in Y and errors in the prior group, the X by SGS-FL is not perfectly sparse in the coefficients of the samples in the unselected groups.

C. Controlling Group Sparsity by Parameter r

Selecting appropriate group sparsity with r is important for the performance of SGS-FL. We tested SGS-FL with different $r \in [0, 1]$ with step size 0.05 and plot the accuracy of the learned X and W by calculating the Pearson correlation with the original ones in Fig. 5. As expected that the group selection changes by steps (Eqn. 3) and the performance of SGS-FL only changes when group selection changes. Thus, SGS-FL performs the same in a certain range of r until reaching an increase or a decrease in the number of selected groups. It is interesting that in the range $r \in [0.45, 0.65]$, X and W are most accurately recovered, and when r is

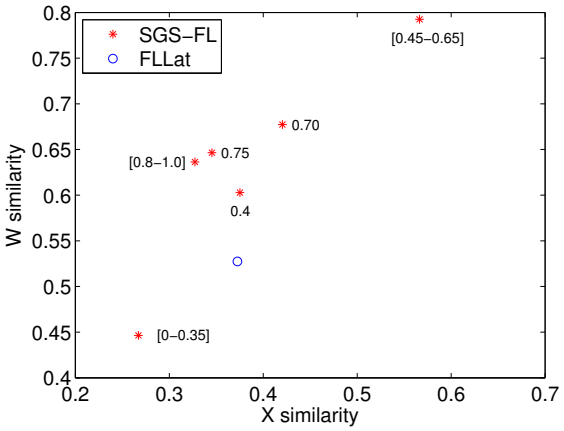


Figure 5. Accuracy of the learned X and W at different group sparsities. Different group sparsity parameter r is tested. The x-axis is the Pearson correlation between the learned X and the original X, and the y-axis is the correlation between the learned W and the original W. Each star represents the accuracy of X and W, labeled by the corresponding r parameter.

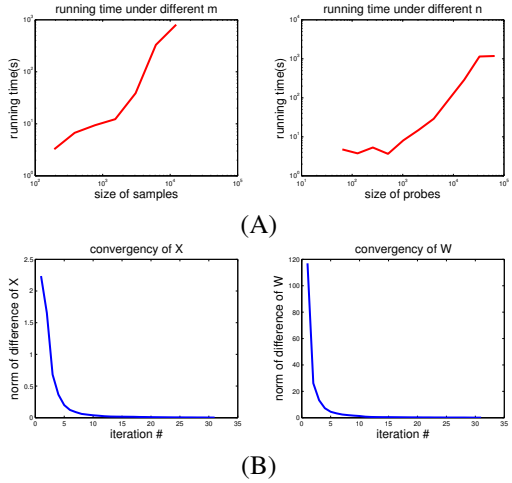


Figure 6. Running time and convergency of SGS-FL. (A) Running time under different m (# of samples) and n (# of probes). (B) Convergency of X and W.

too small or too large, the correlations are lower. This is consistent with our hypothesis: a small r leads to insufficient group selection for the components and a large r may lead to unnecessary group selection and thus an overly dense X that overfits Y. It is clear that in a reasonable range of sparsity, SGS-FL performs well and SGS-FL also performs better than FLLat in most of the choices of r.

D. Scalability and Convergence

In real arrayCGH datasets, the number of probes can be as many as several millions. In Fig. 6, we analyze the running time and the convergency of SGS-FL on simulated datasets of different sample sizes and different numbers of probes. In the left plot of Fig. 6(A), we fixed the number of probes to be 300 and vary the sample size; in the right plot, we fixed the sample size to be 150 and vary the number of probes. In both cases, SGS-FL scales linearly with the log of the sizes. In Fig. 6(B), SGS-FL clearly converges within tens of iterations. The results suggest a good scalability to large datasets by SGS-FL.

V. Experiments on Breast Cancer Data

To directly compare SGS-FL with FLLat, we followed the experiment setup in [8]. SGS-FL and FLLat were applied to chromosome 8 and 17 of a breast cancer arrayCGH data from [13] to identify CNV regions for cancer relevance.

A. Breast Cancer Data

The breast cancer data contains profiles of 44 predominantly advanced primary breast tumors with 241 mapped human genes from chromosome 8 and 382 mapped human genes from chromosome 17. Among the 44 profiles, 5 are in tumor grade 1, 21 in grade 2 and 17 in grade 3. This prior clinical information was used in our model to

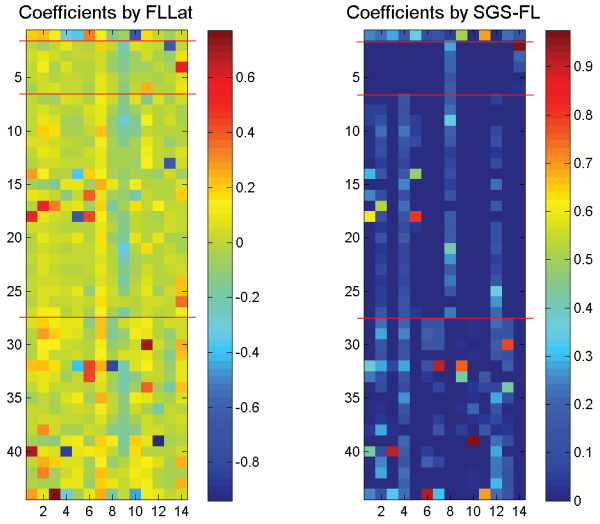


Figure 7. The coefficient matrices learned from the breast cancer data by FLLat and SGS-FL. The sample profile missing the tumor grade information is put in the first row and the other samples are ordered by tumor grade 1-3 from top to bottom. The groups are separated by red horizontal lines. The K columns of X are sorted in descending order by the magnitude of the corresponding latent features (i.e. $\|W_{i,\bullet}\|_2$).

define three groups of samples. There is one additional sample missing the clinical information of tumor grade which was also included in the study. Note that SGS-FL allows additional samples that are not assigned to any group. The number of latent feature K was chosen as the number of principle components that explain at least 80% variation of the data. The hyper-parameters λ_1 and λ_2 were chosen by Bayesian Information Criterion (BIC) suggested by [8]. The r parameter for SGS-FL was set to 0.5 to learn a coefficient matrix with moderate sparsity.

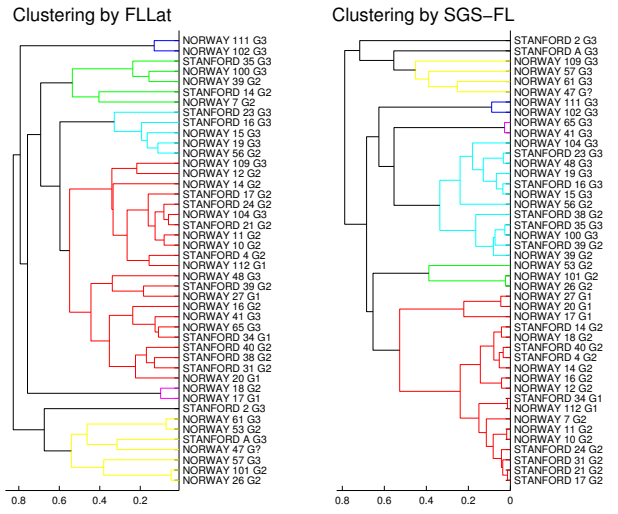


Figure 8. Hierarchical clustering results of breast cancer samples. The samples are labeled by their tumor grade, G1, G2 or G3.

B. Analysis of the Coefficient Matrices

We compare the coefficient matrices learned from FLLat and SGS-FL for chromosome 17 in Fig. 7. The samples are ordered by tumor grade and the three groups are separated by red horizontal lines. In the coefficient matrix learned by FLLat, there is hardly any group structure of samples and the relation between the samples and the latent features seems arbitrary. In other words, there is no subset of samples with similar tumor grade by which a latent feature is shared. The coefficients learned by SGS-FL show clear group structures. When the coefficients are all zeros in one group, it implies that the CNVs identified from the corresponding latent feature are not associated with that group. For example, the first two latent features are only shared by the groups of tumor grade 2 and 3; the third latent feature is only shared by the group of tumor grade 3 and the last latent feature is only shared by the group of tumor grade 1. The submatrix of the last group is denser than those of the first and the second groups, which implies that the samples with tumor grade 3 are sharing more CNVs than the samples in the other two groups.

We also performed hierarchical clustering on the two coefficient matrices. Cosine similarity was used as the similarity metric in the clustering to obtain similar clustering results reported in [8]. The clustering results are shown in Fig. 8. Since the tumor grade information is incorporated in SGS-FL, the generated hierarchical structures are more biologically meaningful. For example, there are three large clusters: the first cluster contains samples with tumor grade 1 and 2, the second cluster contains samples with tumor grade 2 and 3, and the third cluster contains samples with tumor grade 3 except the additional sample missing the tumor grade information. Since tumor grade 2 is an intermediate state between tumor grade 1 and 3, it is reasonable to assume that some samples with tumor grade 2 are more similar to samples with tumor grade 1 and some are more similar to samples with tumor grade 3. The results can be easily explained by the coefficient matrix in Fig. 7. The components can only be shared by samples in group 1 and group 2, or by samples in group 2 and group 3, and never shared by samples in group 1 and group 3. This result strongly support the hypothesis that CNVs correlate with the tumor grade. The clustering result generated by FLLat in Fig. 8 and [8] also showed there are three distinct groups of samples. However, it is not clear why these samples were clustered together since their tumor grades are different.

Table I
CLASSIFICATION ACCURACIES IN LEAVE-ONE-OUT CROSS-VALIDATION WITH BEST RESULTS FOR EACH METHOD BOLD.

	k=1	k=3	k=5	k=7
chromosome 8				
PCA	0.727	**0.795**	**0.795**	0.682
FLLat	**0.659**	**0.659**	0.614	0.636
SGS-FL	0.705	**0.795**	0.750	0.773
chromosome 17				
PCA	0.750	0.795	**0.818**	0.750
FLLat	0.591	0.682	0.659	**0.750**
SGS-FL	**0.773**	0.750	0.705	0.750
Tumor Grade	0.727	0.727	0.727	0.727

C. Sample Classification

To check whether the coefficient matrix X is also consistent with other clinical information, we also designed a binary classification problem of separating samples into two groups with another clinical variable 'Tumor size' (T1&T2 vs. T3&T4). Tumor grades were used as prior group information by SGS-FL and the 'Tumor size' variable is the target variable for classification. We run a leave-one-out cross-validation with k-nearest neighbor (KNN) classifier on the coefficient matrices learned by PCA, FLLat and SGS-FL from chromosome 8 and 17. The number of latent features are fixed to be the number of principle components that explain 80% variance for all the three methods. The classification accuracies by the three methods with different KNN parameters are reported in Tab. I. It is not surprising that PCA achieved the best performance since PCA preserves the most variance of the data without constraints on obtaining interpretable coefficients. Nevertheless, in the table the result by SGS-FL is comparable to PCA and much better than the result by FLLat. It suggests that by using relevant prior information, SGS-FL can obtain both interpretable CNV components and informative coefficients for classification. It is worth noting that if only the tumor grade information is used by KNN in the leave-one-out cross-validation, the accuracy is 0.727 for any choice of k for the KNN classifier. This result further implies that the better classification performance of SGS-FL is not solely due to the relevant prior information in tumor grade.

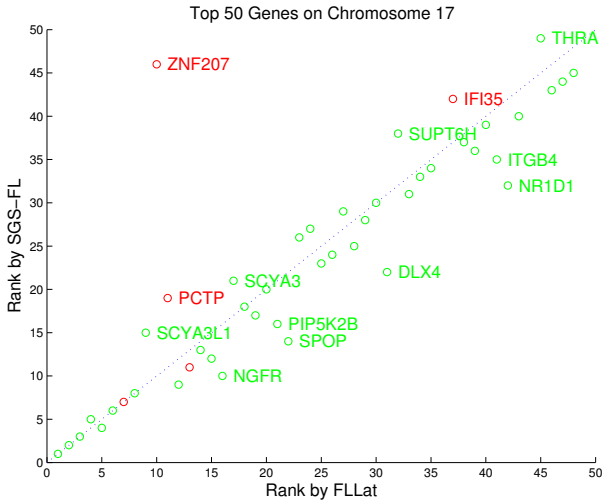


Figure 9. Top ranked 50 genes by FLLat and SGS-FL on chromosome 17. Green denotes 'gain' status and red denotes 'loss' status. Genes with most different ranks by FLLat and SGS-FL are label by their gene symbols.

D. Analysis of CNV Components

We next compared the latent features learned by FLLat and SGS-FL. We ranked the identified probes by the sum of their magnitude in all latent features (i.e. $\sum_k |W_{kj}|$). The probes without gene names were excluded in this analysis. We took the top-50 genes ranked by FLLat and SGS-FL, and plot their ranks in Fig. 9. FLLat and SGS-FL have consensus on the ranks of many of the genes. We focus on the genes which have a difference larger than 3 in the ranks by the two methods. There are several interesting examples. NGFR was demonstrated as a marker to identify myoepithelial cells in preinvasive lesions and myoepithelial differentiation in breast carcinomas [14]. SPOP can mediate the Breast cancer metastasis suppressor 1 (BRMS1) and is important for breast cancer progression [15]. PIP5K2B (PIP4K2B) is a known amplified gene in breast cancer [16]. DLX4 (BP1) negatively regulates BRCA1 in sporadic breast cancer [17]. NR1D1 is a survival factor for breast cancer [18]. ITGB4 is a prognostic marker for breast cancer [19]. All the genes ranked better by SGS-FL seem to be relevant to breast cancer. However, for the genes ranked higher by FLLat such as ZNF207, PCTP and SCYA3L1, there is no literature suggesting associations between the genes and breast cancer. Possibly, these genes might be involved in some frequent CNVs instead of CNVs specific to breast cancer.

Finally, we also compared the ranking of the known cancer genes in Cancer Gene Census[1] on chromosome 8 and 17 in Fig. 10. Overall, most of the known cancer genes were ranked better by SGS-FL. The result implies that the identified CNVs by SGS-FL are more likely to be associated with breast cancer.

[1]http://www.sanger.ac.uk/genetics/CGP/Census/

Table II
RANKING OF THE 33 KNOWN AMPLIFIED GENES IN THE BLADDER CANCER DATA. THE BEST RANK OF EACH GENE IS BOLD.

Chromosome	Gene	Naïve	FLLat	SGS-FL
2	CPSF3	3	**1**	**1**
	ADAM17	3	**1**	**1**
	YWHAQ	3	**1**	**1**
	TAF1B	8	**4**	**4**
	UNQ5830	13	**5**	**5**
	KLF11	**1**	5	5
	RRM2	**1**	8	8
6	CAP2	53	59	**52**
	FAM8A1	**28**	61	53
	NUP153	**28**	61	53
	KIF13A	**28**	62	54
	NHLRC1	79	66	**58**
	AOF1	117	67	**59**
	DEK	117	67	**59**
	IBRDC2	117	67	**59**
	ID4	66	31	**30**
	OACT1	23	**13**	14
	E2F3	10	**1**	**1**
	CDKAL1	**3**	**3**	**3**
	SOX4	20	**16**	**16**
	PRL	**15**	22	22
8	COX6C	65	49	**45**
	POLR2K	69	51	**47**
	SPAG1	69	51	**47**
	RNF19	69	51	**47**
	MGC39715	76	53	**49**
	NCALD	**6**	7	37
	RRM2B	52	**10**	39
	ODF1	138	**63**	71
	KLF10	185	**64**	76
	FLJ45248	162	**65**	77
	ATP6V1C1	46	58	**35**
	BAALC	68	60	**41**

VI. EXPERIMENTS ON BLADDER CANCER DATA

We also applied SGS-FL and FLLat to test a bladder cancer arrayCGH data from [9] to identify CNVs relevant to bladder cancer. This dataset contains 38 urothelial carcinomas with whole-genome tiling resolution array-CGH and high density expression profiling. We still used tumor grade as the prior information to separate samples into 3 groups $\{$G1,G2,G3$\}$ and set $r = 0.5$ for SGS-FL. The parameters k, λ_1 and λ_2 were selected in the same way as in the previous experiment. [9] reported genomic amplifications of 47 genes at regions 2p25, 6p22 and 8q22 in "Additional file 4", so we focused our study on these chromosomes. There are 1938 probes on chromosome 2; 1801 probes on chromosome 6; and 1091 probes on chromosome 8. 33 of the 47 genes are annotated in the dataset. Both FLLat and SGS-FL identified the 33 known amplified genes and ranked them in the top 100 probes. We also compared the methods with the naïve approach which ranks the genes simply based on the sum of their magnitude in the original data (i.e. $\sum_i |Y_{ij}|$). The 33 genes and their corresponding ranks are listed in Tab. II. Compared with FLLat, SGS-FL ranked 16 genes better and 6 genes worse. The result suggests that the prior information in tumor grade helps rank the cancer relevant CNVs higher.

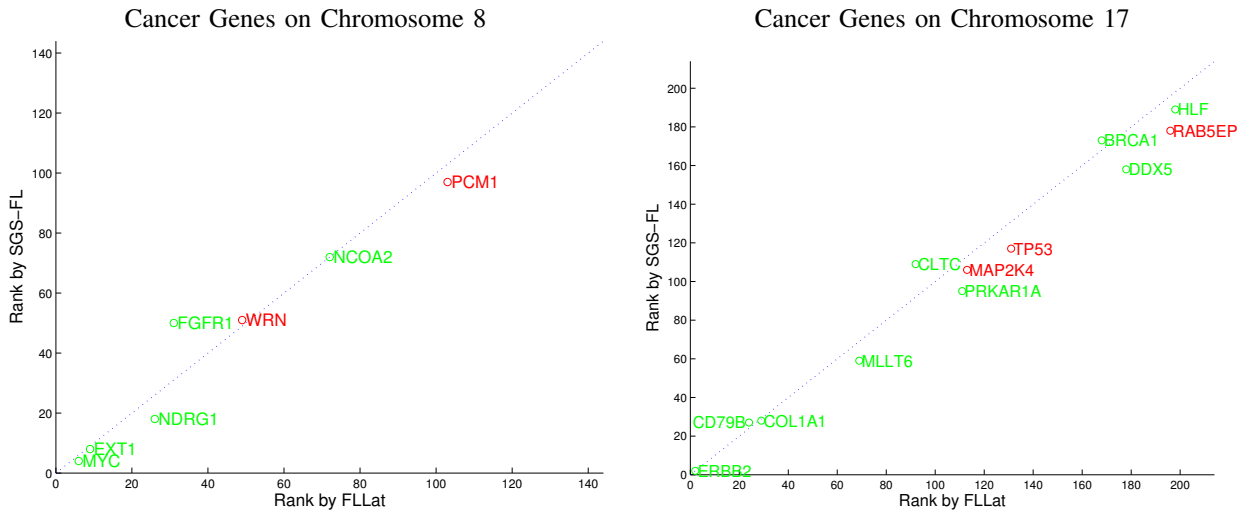


Figure 10. Ranking of known cancer genes on chromosome 8 and 17. Green denotes 'gain' status and red denotes 'loss' status. On average, SGS-FL ranked the cancer genes on chromosome 8 and 17 better than FLLat.

Compared with the naïve method, SGS-FL ranked 25 genes better and 7 genes worse. The result suggests that the learned latent features is more reliable than the original data for identifying cancer relevant CNVs.

VII. Conclusions

In general, discovering CNVs across multiple samples is more accurate than single sample analysis. To analyze multiple samples of probe series, it is important to consider both the similarity and the heterogeneity among the samples. Existing methods such as FLLat ignore the fact that patient samples with different phenotypes show different frequencies and patterns of CNVs in their genotyping. These methods tend to miss the CNVs specific to subsets of samples. To the best of our knowledge, SGS-FL is the first model that considers the prior information on sample groups in CNV identification. SGS-FL constructs a latent feature model to identify CNVs and learn the sample groups sharing the CNVs simultaneously by integrating fused lasso to smooth CNV patterns and adaptive sparse group selection to identify the group specificity of the CNVs. The simulations and experiments on real cancer arrayCGH datasets suggest that with the relevant sample group information, SGS-FL can more accurately identify cancer relevant CNV regions and a more informative representation of CNV data as coefficients on the CNV components.

Acknowledgment

This work is supported by NSF grant #IIS1117153. The authors gratefully thank Christina Leslie and Jieping Ye for helpful discussions.

References

[1] L. Feuk, A. Carson, and S. Scherer, "Structural variation in the human genome," *Nature Reviews Genetics*, vol. 7, no. 2, pp. 85–97, FEB 2006.

[2] R. Redon *et al.*, "Global variation in copy number in the human genome," *Nature*, vol. 444, no. 7118, pp. 444–454, NOV 23 2006.

[3] N. P. Carter, "Methods and strategies for analyzing copy number variation using DNA microarrays," *Nature Genetics*, vol. 39, no. 7, pp. S16–S21, JUL 2007.

[4] N. H. Sykes *et al.*, "Copy number variation and association analysis of SHANK3 as a candidate gene for autism in the IMGSAC collection," *European Journal Of Human Genetics*, vol. 17, no. 10, pp. 1347–1353, OCT 2009.

[5] R. Tibshirani and P. Wang, "Spatial smoothing and hot spot detection for CGH data using the fused lasso," *Biostatistics*, vol. 9, no. 1, pp. 18–29, JAN 2008.

[6] J.-P. Vert and K. Bleakley, "Fast detection of multiple change-points shared by many signals using group lars," in *Advances in Neural Information Processing Systems 23*, J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, Eds., 2010, pp. 2343–2351.

[7] K. Bleakley and J.-P. Vert, "The group fused lasso for multiple change-point detection," 2011.

[8] G. Nowak, T. Hastie, J. R. Pollack, and R. Tibshirani, "A fused lasso latent feature model for analyzing multi-sample aCGH data," *Biostatistics*, vol. 12, no. 4, pp. 776–791, OCT 2011.

[9] M. Heidenblad *et al.*, "Tiling resolution array CGH and high density expression profiling of urothelial carcinomas delineate genomic amplicons and candidate target genes specific for advanced tumors," *BMC Medical Genomics*, vol. 1, JAN 31 2008.

[10] J. Liu, S. Ji, and J. Ye, *SLEP: Sparse Learning with Efficient Projections*, Arizona State University, 2009. [Online]. Available: http://www.public.asu.edu/~jye02/Software/SLEP

[11] J. Kim, R. Monteiro, and H. Park, "Group sparsity in nonnegative matrix factorization," in *Proceedings of the 12th SIAM International Conference on Data Mining*, 2012, p. 851.

[12] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Convex optimization with sparsity-inducing norms."

[13] J. Pollack *et al.*, "Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 20, pp. 12 963–12 968, OCT 1 2002.

[14] J. Reis *et al.*, "Distribution and significance of nerve growth factor receptor (NGFR/p75(NTR)) in normal, benign and malignant breast tissue," *Modern Pathology*, vol. 19, no. 2, pp. 307–319, FEB 2006.

[15] B. Kim *et al.*, "Breast cancer metastasis suppressor 1 (BRMS1) is destabilized by the Cul3-SPOP E3 ubiquitin ligase complex," *Biochemical And Biophysical Research Communications*, vol. 415, no. 4, pp. 720–726, DEC 2 2011.

[16] S. Luoh, N. Venkatesan, and R. Tripathi, "Overexpression of the amplified Pip4k2 beta gene from 17q11-12 in breast cancer cells confers proliferation advantage," *Oncogene*, vol. 23, no. 7, pp. 1354–1363, FEB 19 2004.

[17] B. J. Kluk *et al.*, "BP1, an Isoform of DLX4 Homeoprotein, Negatively Regulates BRCA1 in Sporadic Breast Cancer," *International Journal of Biological Sciences*, vol. 6, no. 5, pp. 513–524, 2010.

[18] A. Kourtidis *et al.*, "An RNA Interference Screen Identifies Metabolic Regulators NR1D1 and PBP as Novel Survival Factors for Breast Cancer Cells with the ERBB2 Signature," *Cancer Research*, vol. 70, no. 5, pp. 1783–1792, MAR 1 2010.

[19] A. Brendle *et al.*, "Polymorphisms in predicted microRNA-binding sites in integrin genes and breast cancer: ITGB4 as prognostic marker," *Carcinogenesis*, vol. 29, no. 7, pp. 1394–1399, JUL 2008.