University of Wisconsin Department of Biostatistics and Medical Informatics

Technical Report # 219

Title: Sparse Integrative Clustering Of Multiple Omics Data Sets

Authors: Ronglai Shen, Sijian Wang and Qianxing Mo

Date: June 2011

SPARSE INTEGRATIVE CLUSTERING OF MULTIPLE OMICS DATA SETS

BY RONGLAI SHEN [§], SIJIAN WANG [¶] AND QIANXING MO [§]

Memorial Sloan-Kettering Cancer Center[§] *University of Wisconsin, Madison*[¶]

Abstract. High resolution microarrays and second-generation sequencing platforms are powerful tools to investigate genome-wide alterations in DNA copy number, methylation, and gene expression associated with a disease. An integrated genomic profiling approach measuring these omics data types simultaneously in the same set of biological samples would render an integrated data resolution that would not be available with any single data type. In the context of integrative genomics, we propose a statistical framework that aims to address clustering, data integration, dimension reduction, and variable selection in a unified and computationally efficient way. We derive sparse integrative clustering solutions by incorporating regularization methods including the Lasso, Elasticnet, and Fused Lasso. The proposed method is applied to integrate genomic, epigenomic, and transcriptomic data for subtype analysis in breast cancer.

1. Introduction.

Clustering analysis is an unsupervised learning method that aims to group data into distinct clusters based on a certain measure of similarity among the data points. Clustering analysis has many applications in a wide variety of fields including pattern recognition, image processing and bioinformatics. In gene expression microarray studies, clustering cancer samples based on their gene expression profile has revealed molecular subgroups associated with histopathological categories, drug response, and patient survival differences (Alizadeh *et al.*, 2000; Perou *et al.*, 1999; Sorlie *et al.*, 2001).

In the past few years, *integrative genomic studies* are emerging at a fast pace where in addition to gene expression data, genome-wide data sets capturing somatic mutation patterns, DNA copy number alterations, DNA

[*] Some comment

[†] First supporter of the project

[‡] Second supporter of the project

methylation changes are simultaneously obtained in the same biological samples. A fundamental challenge in translating cancer genomic findings into clinical application lies in the ability to find “driver” genetic and genomic alterations that contribute to tumor initiation, progression, and metastasis (Chin and Gray, 2008; Simon, 2010). As integrated genomic studies have emerged, it has become increasingly clear that true oncogenic mechanisms are more visible when combining evidence across patterns of alterations in DNA copy number, methylation, gene expression and mutational profiles (TCGA, 2008; Chitale *et al.*, 2009; Cerami *et al.*, 2010; Vaske *et al.*, 2010). Integrative analysis of multiple “omic” data types can help the search for “drivers” by uncovering genomic features that tend to be dysregulated by multiple mechanisms (Chin and Gray, 2008).

A classic example is the tumor-suppressor protein INK4A (encoded by CDKN2A), which can be inactivated through homozygous loss (copy number), epigenetic silencing (by promoter methylation), or loss-of-function mutations in the protein (Sharpless, 2005). Analogously, the HER2 oncogene can be activated through DNA amplification and mRNA overexpression which we will discuss further in our motivating example. Individually, none of the genomic-wide data type alone can completely capture the complexity of the cancer genome or fully explain the underlying disease mechanism. Collectively, however, true oncogenic mechanisms may emerge as a result of joint analysis of multiple genomic data types.

In this paper, we focus on class discovery problem given multiple omics data sets (multidimensional data) for tumor subtype discovery. In the past years, molecular subtype discovery based on gene expression microarray data alone has been widely applied in cancer research. Nevertheless, the clinical and therapeutic implications for these existing molecular subtypes of cancer are largely unknown. A confounding factor is that expression changes may be related to cellular activities independent of tumorigenesis, and therefore leading to subtypes that may not be directly relevant for diagnostic and prognostic purposes. By contrast, joint analysis of multiple omics data types offer a new paradigm to gain additional insights, and to help distinguish cancer driving factors that define biologically and clinically relevant subtypes of the disease by uncovering genomic markers with coordinated patterns of alteration in the cancer genome, epigenome, and transcriptome.

Large-scale cancer genome projects including the Cancer genome Atlas (TCGA) project and the International Cancer Genome Consortium (ICGC) project are generating an unprecedented amount of multidimensional data sets using high resolution microarray and next-generation sequencing platforms. With the accumulating wealth of multidimensional data, new integrative analysis methods are emerging in this field. Van Wieringen and Van de Wiel (2009) proposed a nonparametric testing procedure for DNA copy number induced differential mRNA gene expression. Peng *et al.* (2010) and Vaske *et al.* (2010) considered pathway and network analysis using multiple genomic data sources. A number of others (Waaijenborg, Verselewel de Witt Hamer and Zwinderman, 2008; Parkhomenko, Tritchler and Beyene, 2009; Le Cao, Martin and Robert-Granie, 2009; Witten, Tibshirani and Hastie, 2009; Witten and Tibshirani, 2009; Soneson *et al.*, 2010) suggested using canonical correlation analysis (CCA) to quantify the correlation between two data sets (e.g., gene expression and copy number data). Most of these previous work focused on integrating copy number and gene expression data, and none of these methods were specifically designed for tumor subtype analysis.

Another important problem in high dimensional genomic data analysis is variable selection, which crucially affects model interpretation, computational complexity, and statistical accuracy. Penalized likelihood-based methods have drawn a lot of attention in recent literature, including Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001), Elastic Net (Zou and Hastie, 2005), Fused Lasso (Tibshirani *et al.*, 2005), group lasso (Yuan and Lin, 2006), CAP (Zhao, Rocha and Yu, 2009), MCP (Zhang, 2010) among others. Bayesian variable selection strategies (George and McCulloch, 1993; George and Foster, 2000; Park and Casella, 2008) have also become popular in many applications. In clustering, penalized method has been proposed for both distance-based (Witten and Tibshirani, 2010) and model-based clustering analysis (Pan and Shen, 2007; Xie and Pan, 2008; Wang and Zhu, 2008).

In this paper, we propose a general framework called iCluster for simultaneous data integration, variable selection, and dimension reduction for integrative subtype analysis of any number of omics data sets. The proposed method allows *de novo* identification of important cancer driving factors dysregulated by multiple mechanisms using joint latent variable models. The penalized maximum likelihood method is used for a joint feature selection to reveal genomic alterations closely associated with the cancer driving factors. It is worth pointing out that not all genomic alterations can be interpreted in a gene-centric fashion (such as microRNA or alterations in intronic/intergenic regions). To facilitate the discussion of variable selection in the Method Section, we use genomic feature as a general term to refer to protein-coding genes as well as non-coding genetic and genomic elements depending on the platform and data type. In Shen, Olshen and Ladanyi (2009), we give a brief description of the method, focusing on the biological findings from two different applications. In the current paper, we give a broader and more thorough and rigorous statistical treatment. Computationally, we derive a unified algorithm for a range of regularized minimization problems and show that this new implementation scales well for increasing data dimension.

The rest of the article is organized as follows. In Section 2, we present a motivating example for our proposed iCluster modeling approach. In Section 3, we describe the method. In Section 4, we propose a unified algorithm that efficiently solves a range of regularized minimization problems for model fitting in the iCluster framework. We also discuss the selection of number of clusters based on a resampling-based approach and a uniform design for the tuning parameter selection, which is more efficient than the usual grid search based strategy. In Section 5 and 6, we assess the performance of the proposed method using simulated and real data sets. We conclude the paper with a brief summary in Section 7.

2. A Motivating Example.

The work in this paper was motivated by analyzing the Pollack *et al.* (2002) data set with *parallel* microarray measurements of DNA copy number and mRNA expression in 37 primary breast cancer and 4 breast cancer cell line samples. The study was among the first to reveal on a global scale a high degree to which variation in gene copy number contributes to variation in gene expression in tumor cells. Specifically, it was estimated that 62% of highly amplified genes demonstrate moderately or highly elevated expression level, and about 12% of all variation in gene expression changes can be directly attributed to DNA copy number aberrations.

Separate clustering analysis is performed to reveal biologically meaningful subgroups in DNA copy number and in mRNA expression level respectively. A heatmap of the genomic features on chromosome 17 is plotted in Figure 1. Rows are genomic features ordered by chromosomal position and columns are samples ordered by clustering. There are two main subclasses in the 41 samples: the cell line subclass (samples labeled in red) and the HER2 tumor subclass (samples labeled in green). These subclasses are not discriminated well from separate hierarchical clustering analyses shown in Figure 1A.

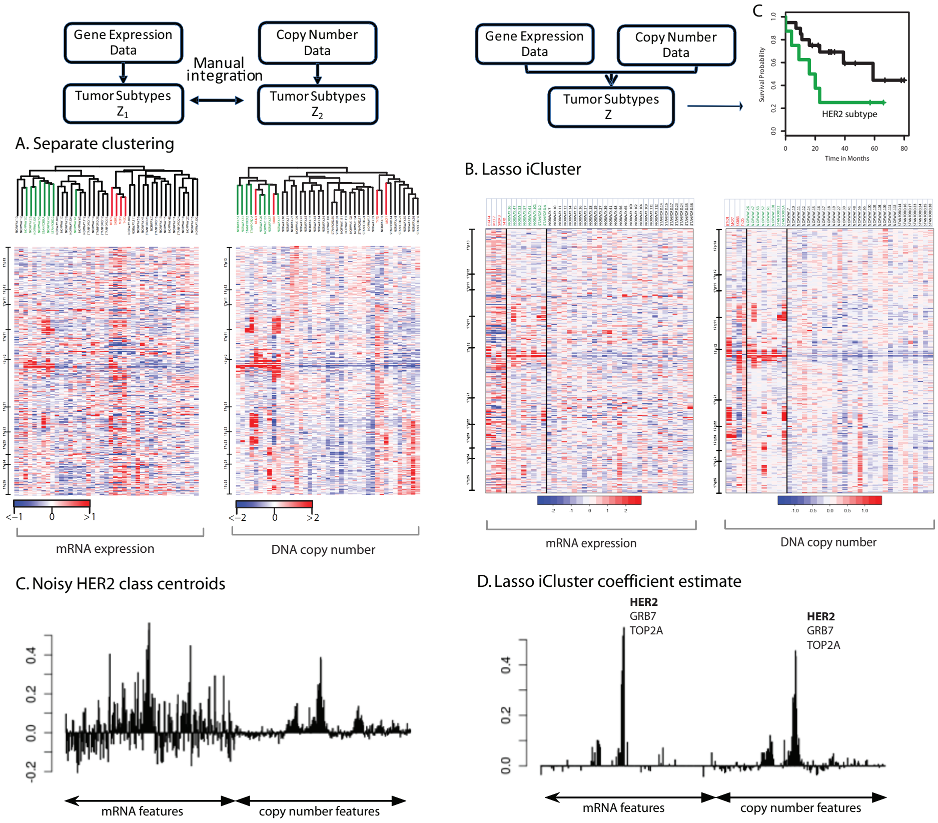

FIG 1. *A motivating example using the Pollack data set to demonstrate that a joint analysis using the Lasso iCluster outperforms the separate clustering approach in subtype analysis given DNA copy number and mRNA expression data. The top panel includes heatmaps with features (row) ordered by chromosomal position and samples (column) arranged by separate hierarchical clustering (A) and by Lasso iCluster (B). Two distinct subclasses exist: the cell line subclass (samples labeled in red) and the HER2 tumor subclass (samples labeled in green). In the bottom panel (C and D), the Lasso iCluster coefficient estimates clearly reveal HER2-subtype specific genes, superior to the standard cluster centroid estimates.*

Separate clustering followed by manual integration remains the most frequently applied approach to analyze multiple omics data sets in the current literature for its simplicity and the lack of a truly integrative approach. However, Figure 1A clearly shows its lack of consistency in cluster assignment and poor correlation of the outcome with biological and clinical annotation. As we will illustrate in the simulation study in Section 6, separate clustering can fail drastically in estimating the true number of clusters, classifying samples to the correct clusters, and selecting cluster-associated features. Several limitations of this common approach are responsible for its poor performance. First, correlation between data sets is not utilized to inform the clustering analysis. As discussed earlier, correlation plays a key role for identifying "driver" features of biological importance. In addition, separate analysis of *paired* genomic data sets is an inefficient use of the available information. Second, it is not straightforward to integrate the multiple sets of cluster assignments that are data-type dependent without extensive prior information on cancer biology. Finally, the common approach includes all genomic features regardless of their relevance to clustering.

Our method aims to overcome these obstacles by formulating a joint analysis across multiple omics data sets. The heatmap in Figure 1B demonstrates the superiority of our working model in correctly identifying the subgroups (vertically divided by solid black lines). From left to right, cluster 1 (samples labeled in red) corresponds to the breast cancer cell line subgroup, distinguishing cell line samples from tumor samples. Cluster 2 corresponds to the *HER2* tumor subtype (samples labeled in green), showing concordant amplification in the DNA and overexpression in mRNA at the *HER2* locus (chr 17q12). This subtype is associated with poor survival as shown in Figure 1C. iCluster 3 (samples labeled in black) did not show any distinct patterns, though a pattern may have emerged if there were additional data types such as DNA methylation.

The motivation for sparseness in the estimated clustering coefficients (the interpretation of the coefficients will be explained in Section 3) is illustrated by Figure 1D. It clearly reveals the *HER2*-subtype specific genes (including *HER2, GRB7, TOP2A*). By contrast, the standard cluster centroid estimation is flooded with noise (Figure 1C), revealing an inherent problem with clustering methods without regularization.

3. Method.

Our method is mainly motivated by identifying a set of driving factors that define biologically and clinically relevant subtypes of the disease. In the motivating example, the *HER2* breast tumor subtype was revealed through the identification of the driving factor: *HER2* activation. In general, if we have knowledge on the driving factors (such as *HER2*) in a certain cancer and observe their values (degree of *HER2* activation through amplification and overexpression) for each tumor, we can model the alterations across multiple omics data spaces as functions of the driving factors for effective data integration and dimension reduction.

In the general class discovery setting, however, these driving factors are unobserved due to the complexity of the cancer genome and limited knowledge on major cancer driving genes. This motivates us to consider a latent variable modeling framework. To be specific, we assume that each tumor can be represented by a latent random variable, and the values of genomic feature sets (DNA copy number, mRNA expression, methylation) are determined by certain functions based on the corresponding latent variable. Furthermore, through the shared latent variables, our method is able to induce correlation among different genomic data types, which may play a critical role in identifying such cancer genes of biological and clinical importance. Figure 2 graphically illustrates this latent variable modeling approach. The details of the modeling approach is illustrated as follows.

Suppose $t = 1, \cdots, m$ different genome-scale data types (DNA copy number, methylation, mRNA expression, etc.) are obtained in $i = 1, \cdots, n$ tumors. Let $\boldsymbol{x}_{it} = (x_{i1t}, \cdots, x_{ip_t t})'$ denote a p_t-dimensional data vector. Each element x_{ijt}, $j = 1, \cdots, p_t$, represents the observation associated with the jth genomic feature in data type t measured in tumor i.

In matrix form, $\boldsymbol{X}_t = (\boldsymbol{x}_{1t}, \cdots, \boldsymbol{x}_{nt})$, $t = 1, \cdots, m$ is the tth data matrix of dimension $p_t \times n$. We assume that the observed genomic data sets are each properly centered to have zero means. Now we further assume there are K subtypes jointly characterize the maximum variation across the m data types (in Section 3.1, we will discuss how to determine K). Let $\mathbf{z}_i$ denote the latent variable of length $K-1$ for the ith tumor. The reason we assume the length of $\boldsymbol{z}_i$ is $K-1$ is that there is at most $K-1$ dimension to separate data points from K clusters (Hastie, Tibshirani and Friedman, 2009). We further assume $\boldsymbol{z}_i \overset{i.i.d.}{\sim} MVN(\mathbf{0}, \boldsymbol{I})$. The reason to assume mean zero is that the variable are centered before the analysis, so the marginal means are zero. The reason to assume an identity covariance matrix is because of the identifiability between $\boldsymbol{w}_{jt}$ and $\boldsymbol{z}_i$ in the following equation (1). Finally, we relate the multidimensional genomic data vectors to the latent factors through a linear function using the following iCluster model

$$(1) \quad x_{ijt} = \boldsymbol{w}'_{jt}\boldsymbol{z}_i + \boldsymbol{\epsilon}_{ijt}, \quad i = 1, \cdots, n; \; j = 1, \cdots p_t; \; t = 1, \cdots, m,$$

where $\boldsymbol{w}_{jt}$ is the length-$(K-1)$ coefficient vector associated with the jth feature in the tth data type, and $\boldsymbol{W}_t = (\boldsymbol{w}_{1t}, \cdots, \boldsymbol{w}_{p_t t})'$ is the corresponding coefficient matrix of dimension $p_t \times (K-1)$. Finally, $\boldsymbol{\epsilon}_{it}$ denotes the error term with mean zero and covariance matrix $\boldsymbol{\Psi}_t = \text{diag}(\sigma^2_{1t}, \cdots, \sigma^2_{p_t t})$, representing the residual variance. Marginally, $\boldsymbol{x}_{it} \sim N(0, \boldsymbol{\Sigma}_t)$ with $\boldsymbol{\Sigma}_t = \boldsymbol{W}_t\boldsymbol{W}_t' + \boldsymbol{\Psi}_t$.

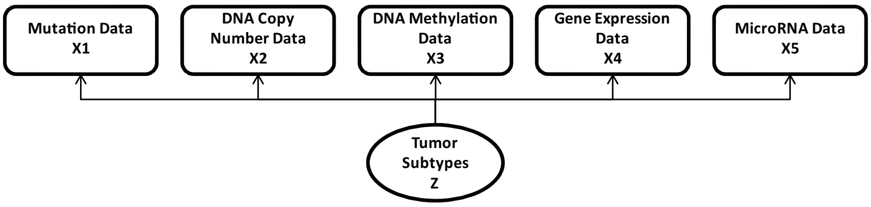

FIG 2. *A novel concept for integrative clustering.*

The sparsity of $\boldsymbol{W}$ directly impacts the interpretability of the latent factors. If $w_{kjt} = 0$, then the kth latent variable has no effect on the jth feature in the tth data type. If $w_{kjt} = 0\,\forall k$, then the genomic feature x_{jt} provides no information for clustering and will be removed from the model. In order to have the desired sparsity in estimated $\hat{\boldsymbol{w}}_{jt}$, we consider the following regularized log-likelihood:

$$\ell_{c,p}(\{\boldsymbol{W}_t\}_{t=1}^m, \{\boldsymbol{\Psi}\}_{t=1}^m) = \ell_c - \sum_{t=1}^m J_{\lambda_t}(\boldsymbol{W}_t), \tag{2}$$

where ℓ_c is the complete-date log-likelihood function of the following form

$$\ell_c \propto -\frac{n}{2}\sum_{t=1}^m \log|\boldsymbol{\Psi}_t| - \frac{1}{2}\sum_{t=1}^m\sum_{i=1}^n(\boldsymbol{x}_{it} - \boldsymbol{W}_t\boldsymbol{z}_i)'\boldsymbol{\Psi_t}^{-1}(\boldsymbol{x}_{it} - \boldsymbol{W}_t\boldsymbol{z}_i) - \frac{1}{2}\sum_{i=1}^n \boldsymbol{z}_i'\boldsymbol{z}_i, \tag{3}$$

In (2), ℓ_c controls the fitness of the model, and $J_{\lambda_t}(\boldsymbol{W}_t)$ is a penalty function which controls the complexity of the model; and λ_t's are non-negative tuning parameters which determines the balance between the two.

In the literature, there are many penalty function being proposed for the purpose of variable selection. We first consider the ℓ_1-penalty (Tibshirani, 1996) that takes the form

$$J_{\lambda_t}(\boldsymbol{W}_t) = \lambda_t \sum_{k=1}^{K-1}\sum_{j=1}^{p_t}|w_{kjt}|. \tag{4}$$

The ℓ_1-penalty continuously shrinks the coefficients toward zero and thereby benefits from the substantial decrease in the variance of the coefficients. Owing to the singularity of ℓ_1-penalty at the origin point ($w_{kjt} = 0$), some estimated $\hat{w}_{kjt}$ will be *exactly* zero. The degree of sparseness is controlled by the tuning parameter λ_t. If the entire row-vector $\boldsymbol{w}_{tj}$ are shrunken to zero, the corresponding genomic feature will be excluded from the model, rendering a sparse feature space.

To account for the genomic ordering in DNA copy number data, we consider the fused lasso penalty (Tibshirani *et al.*, 2005), which takes the following form

$$J(\boldsymbol{W}_t) = \lambda_{1t} \sum_{k}^{K-1} \sum_{j=1}^{p_t} |w_{kjt}| + \lambda_{2t} \sum_{k=1}^{K-1} \sum_{j=2}^{p_t} |w_{kjt} - w_{k(j-1)t}|, \tag{5}$$

where λ_{1t} and λ_{2t} are two non-negative tuning parameters. The first penalty encourages sparseness while the second encourages smoothness along index j. The Fused Lasso penalty is particularly suitable for DNA copy number data where contiguous regions of a chromosome tend to be altered in the same fashion. Tibshirani and Wang (2008) applied a Fused Lasso Signal Approximator (FLSA) for copy number segmentation.

We also implemented the elastic net penalty (Zou and Hastie, 2005), which takes the form

$$J(\boldsymbol{W}_t) = \lambda_{1t} \sum_{k=1}^{K-1} \sum_{j=1}^{p_t} |w_{kjt}| + \lambda_{2t} \sum_{k=1}^{K-1} \sum_{j=1}^{p_t} w_{kjt}^2, \tag{6}$$

where λ_{1t} and λ_{2t} are two non-negative tuning parameters. Zou and Hastie (2005) showed that the elastic net penalty tends to select or remove highly correlated predictors together in linear regression setting by enforcing their estimated coefficients to be similar. In our experience, the elastic net penalty tends to be more numerically stable than lasso penalty in our model.

Figure 3 shows the effectiveness of sparse iCluster using a simulated pair of data sets. There is a subgroup of 50 synthetic subjects jointly characterized by features 1 to 20 in data 1 and features 101 to 120 in data 2. The pair of coefficient vectors $(\boldsymbol{w}_1, \boldsymbol{w}_2)$ estimated from Lasso, Enet, and Fused Lasso iCluster are plotted to contrast the noisy cluster centroids estimated separately in data 1 (left) and in data 2 (right) in the top panel of Figure 3. The cross-validated RI is 0.83, 0.89, and 0.87 for Lasso, Enet, and Fused Lasso iCluster, respectively. In the next Section, we will discuss a unified algorithm used to solve these regularized minimization problems for estimation in the iCluster framework. A systematic simulation study will follow in Section 5.

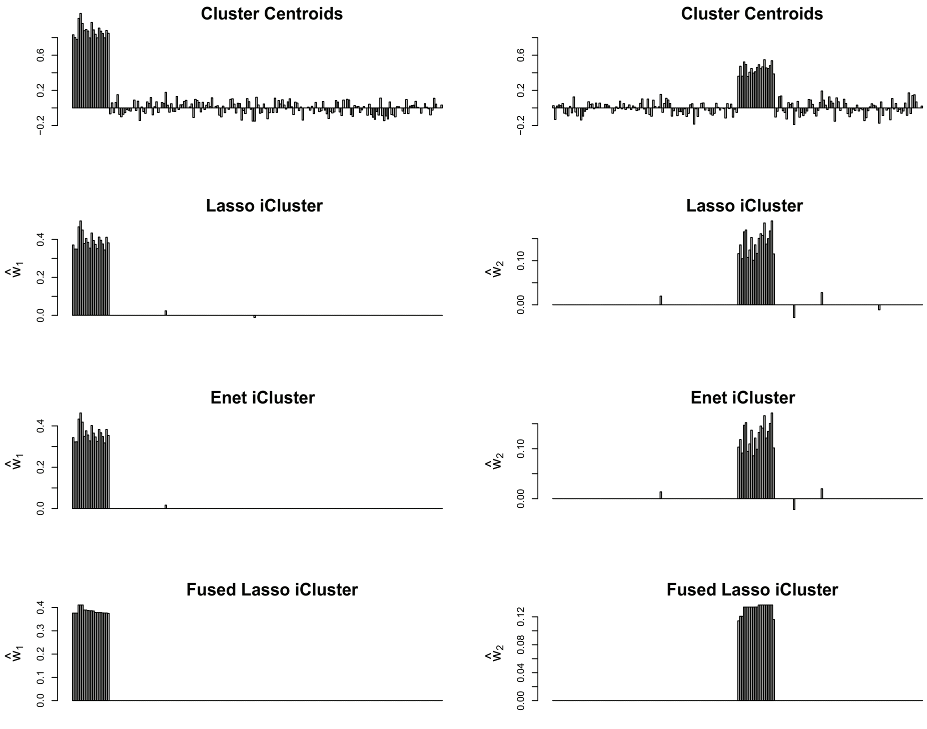

FIG 3. *A simulated pair of data sets each with 100 subjects ($n = 100$) and 200 features ($p_t = 200$, $t = 1, 2$), and 2 subgroups ($K = 2$). Top panel plots the cluster centroids in data 1 (left) and in data 2 (right). Estimated sparse iCluster coefficients are plotted below. The penalty parameters in sparse iCluster are determined by the reproducibility index (RI) in a 10-fold cross-validation.*

4. Algorithm. We now discuss an EM algorithm (Dempster, Laird and Rubin, 1977) to maximize equation (2) for parameter estimation in the sparse iCluster model. The algorithm iterates between an expectation step (E-step) and a penalized maximization step (M-step). When convergence is reached, cluster membership will be assigned for each tumor based on the posterior mean of the latent variable $\boldsymbol{z}_i$. Specifically, in the E-step, we compute the quantity

$$E\left[\ell_{c,p}\middle|\{\boldsymbol{X}_t\}_{t=1}^m, \{\boldsymbol{W}_t^{(q)}\}_{t=1}^m, \{\boldsymbol{\Psi}_t^{(q)}\}_{t=1}^m\right].$$

In the M-step, for $t = 1, \cdots, m$, we obtain the penalized estimates by (7)

$$\hat{\boldsymbol{W}}_t \leftarrow \underset{\boldsymbol{W}_t}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{n} E\Big[(\boldsymbol{x}_{it} - \boldsymbol{W}_t \boldsymbol{z}_i)' \boldsymbol{\Psi}_t^{-1} (\boldsymbol{x}_{it} - \boldsymbol{W}_t \boldsymbol{z}_i) \Big| \boldsymbol{W}_t^{(q)}, \boldsymbol{\Psi}_t^{(q)}\Big] + J_{\lambda_t}(\boldsymbol{W}_t).$$

By setting the first derivative of the minimizing function in (7) with respect to $\boldsymbol{W}$ to be zero and assume the order of derivative and expectation can be exchanged, we have

$$\sum_{i=1}^{n} \boldsymbol{x}_{it} E\big[\boldsymbol{z}_i' \big| \boldsymbol{x}_{it}, \boldsymbol{W}_t^{(q)}, \boldsymbol{\Psi}_t^{(q)}\big] - \boldsymbol{W}_t \sum_{i=1}^{n} E\big[\boldsymbol{z}_i \boldsymbol{z}_i' \big| \boldsymbol{x}_{it}, \boldsymbol{W}_t^{(q)}, \boldsymbol{\Psi}_t^{(q)}\big] - \boldsymbol{\Psi}_t \frac{\partial}{\partial \boldsymbol{W}_t} J(\boldsymbol{W}_t) = 0.$$

For easy of presentation, we omit the data type index t (as the following procedure applies the same for all t) and denote

$$\boldsymbol{C} = \sum_{i=1}^{n} \boldsymbol{x}_{it} E\big[\boldsymbol{z}_i' \big| \boldsymbol{x}_{it}, \boldsymbol{W}_t^{(q)}, \boldsymbol{\Psi}_t^{(q)}\big], \text{and}\, \boldsymbol{Q} = \sum_{i=1}^{n} E\big[\boldsymbol{z}_i \boldsymbol{z}_i' \big| \boldsymbol{x}_{it}, \boldsymbol{W}_t^{(q)}, \boldsymbol{\Psi}_t^{(q)}\big].$$

Then the above becomes

$$\boldsymbol{\Psi} \frac{\partial}{\partial \boldsymbol{W}} J(\boldsymbol{W}) + \boldsymbol{W}\boldsymbol{Q} = \boldsymbol{C}.$$

Due to the diagonal structure of $\boldsymbol{\Psi}$ and "non-coupling" structure of the penalty term, the estimation of $\boldsymbol{w}_j$'s are separable. In other words, we can consider the following p estimating equations:

$$\sigma_j^2 \frac{\partial}{\partial \boldsymbol{w}_j'} J(\boldsymbol{w}_j) + \boldsymbol{w}_j' \boldsymbol{Q} = \boldsymbol{C}_j, \ j = 1, \ldots, p,$$

where $\boldsymbol{C}_j$ is the jth row of $\boldsymbol{C}$. Or equivalently,

$$\sigma_j^2 \frac{\partial}{\partial \boldsymbol{w}_j} J(\boldsymbol{w}_j) + \boldsymbol{Q}' \boldsymbol{w}_j = \boldsymbol{C}_j', \ j = 1, \ldots, p.$$

However, the nondifferentiability of the objective function involving ℓ_1 terms makes the minimization difficult. Following the local quadratic approximation method in Fan and Li (2001), using the fact $|\alpha| = \alpha^2/|\alpha|$ when $\alpha \neq 0$, we consider the following quadratic approximation to the ℓ_1 term:

$$J_{\lambda_t}(\boldsymbol{W}_t)^{lasso} \approx \lambda \sum_{k=1}^{K-1} \sum_{j=1}^{p_t} \frac{w_{kjt}^2}{|\hat{w}_{kjt}^{(q)}|}. \tag{8}$$

and obtain the solution for (7) under Lasso penalty by iteratively computing the following ridge type regression

(9)
$$\boldsymbol{w}_{jt} = \Big(\sum_{i=1}^{n} E\big[\boldsymbol{z}_i\boldsymbol{z}_i'|\boldsymbol{x_{it}}, \boldsymbol{W}_t^{(q)}, \boldsymbol{\Psi}_t^{(q)}\big] + \boldsymbol{A}_j\Big)^{-1} \sum_{i=1}^{n} x_{itj} E\big[\boldsymbol{z}_i'|\boldsymbol{x_i}, \boldsymbol{W}_t^{(q)}, \boldsymbol{\Psi}_t^{(q)}\big],$$

where $\boldsymbol{A}_j = 2\sigma_j^2\lambda\text{diag}\{1/|w_{t1j}^{(q)}|, \dots, 1/|w_{t(K-1)j}^{(q)}|\}$ for $j = 1, \cdots, p_t$. Unlike the standard Ridge regression estimates typically require the inversion of $p \times p$ matrix, computing (9) only requires the inversion of a $(K-1) \times (K-1)$ matrix in the latent subspace.

Similarly we consider a quadratic approximation to the ℓ_1 term in the Elasticnet penalty:

$$J_{\lambda_t}(\boldsymbol{W}_t)^{enet} \approx \lambda_{1t} \sum_{k=1}^{K-1} \sum_{j=1}^{p_t} \frac{w_{kjt}^2}{|\hat{w}_{kjt}^{(q)}|} + \lambda_{2t} \sum_{k=1}^{K-1} \sum_{j=1}^{p} w_{kjt}^2, \tag{10}$$

and obtain the solution for (7) by iteratively computing a Ridge regression similar to (9) but with $\boldsymbol{A}_j = 2\sigma_j^2\Big(\lambda_{1t}\text{diag}\{1/|\hat{w}_{1j}^{(q)}|, \dots, 1/|\hat{w}_{(K-1)j}^{(q)}|\} + \lambda_{2t}\boldsymbol{I}\Big)$.

For Fused Lasso penalty,

$$J_{\lambda_t}(\boldsymbol{W}_t)^{fl} \approx \lambda_{1t} \sum_{k=1}^{K-1} \sum_{j=1}^{p} \frac{w_{kjt}^2}{|\hat{w}_{kjt}^{(q)}|} + \lambda_{2t} \sum_{k=1}^{K-1} \sum_{j=2}^{p} \frac{(w_{kjt} - w_{k(j-1)t})^2}{|\hat{w}_{kjt}^{(q)} - \hat{w}_{k(j-1)t}^{(q)}|}. \tag{11}$$

In the Fused Lasso scenario, the parameters are coupled together, and the estimation of $\boldsymbol{w}_{tj}$ are no longer separable. However, we circumvent the problem by expressing the estimating equation in terms of $\boldsymbol{w}_t = \text{vec}(\boldsymbol{W}_t')$, a column vector of dimension $p_t(K-1)$ by stacking the columns of $\boldsymbol{W}_t'$. Then (11) can be expressed in the following form

$$\begin{aligned} J_{\lambda_t}(\boldsymbol{w}_t)^{fl} &\approx \lambda_{1t} \sum_{k=1}^{K-1} \sum_{j=1}^{p} \frac{w_{kjt}^2}{|\hat{w}_{kjt}^{(q)}|} + \lambda_{2t} \sum_{k=1}^{K-1} \sum_{j=2}^{p} \frac{(w_{kjt} - w_{k(j-1)t})^2}{|\hat{w}_{kjt}^{(q)} - \hat{w}_{k(j-1)t}^{(q)}|}. \\ &= \lambda_{1t}\boldsymbol{w}_t'\boldsymbol{A}\boldsymbol{w}_t + \lambda_{2t}\boldsymbol{w}_t'\boldsymbol{L}\boldsymbol{w}_t, \end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{A} &= \mathrm{diag}\Big\{1/|\hat{w}_1^{(q)}|, \ldots, 1/|\hat{w}_s^{(q)}|\Big\} \ (\ s = p_t \cdot (K-1)), \\
\boldsymbol{L} &= \boldsymbol{D} - \boldsymbol{M}, \\
\boldsymbol{M} &= \begin{cases} 1/|\hat{w}_i^{(q)} - \hat{w}_j^{(q)}|, & |i-j| = K-1 \\ 0, & \text{otherwise.} \end{cases} \ (s \times s \text{ dimension}), \\
\boldsymbol{D} &= \mathrm{diag}\{D_1, \ldots, D_s\} \text{ with } D_j \text{ be the summation of the } j \text{ th row of } \boldsymbol{M}.
\end{aligned}$$

The corresponding representation of estimating equation is

$$\frac{\partial}{\partial \boldsymbol{w}} J(\boldsymbol{w}) + \tilde{\boldsymbol{Q}}\boldsymbol{w} = \tilde{\boldsymbol{C}}, \tag{12}$$

where

$$\tilde{\boldsymbol{Q}} = \begin{pmatrix} \sigma_1^{-2}\boldsymbol{Q} & & \\ & \ddots & \\ & & \sigma_{p_t}^{-2}\boldsymbol{Q} \end{pmatrix}, \quad \tilde{\boldsymbol{C}} = \begin{pmatrix} \sigma_1^{-2}\boldsymbol{C}_1' \\ \vdots \\ \sigma_{p_t}^{-2}\boldsymbol{C}_{p_t}' \end{pmatrix}. \tag{13}$$

The solution for (7) under the Fused Lasso penalty is then computed by iteratively computing

$$\boldsymbol{w}_t = \Big(\tilde{\boldsymbol{Q}} + 2\lambda_{1t}\boldsymbol{A} + 2\lambda_{2t}\boldsymbol{L}\Big)^{-1}\tilde{\boldsymbol{C}}. \tag{14}$$

In summary, by the quadratic approximation, we are able to solve sparse iCluster with a completely flexible use of Lasso, Elasticnet, and Fused Lasso all in a unified and computationally efficient fashion.

Finally for $t = 1, \cdots, m$, we update $\boldsymbol{\Psi}$ in the M-step as follows

$$\hat{\boldsymbol{\Psi}}_t = \frac{1}{n}\mathrm{diag}(\boldsymbol{X}_t\boldsymbol{X}_t' - \boldsymbol{W}_t^{(q)}E\big[\boldsymbol{Z}|\{\boldsymbol{X}_t\}_{t=1}^m, \{\boldsymbol{W}_t^{(q)}\}_{t=1}^m, \{\boldsymbol{\Psi}_t^{(q)}\}_{t=1}^m\big]\boldsymbol{X}_t'). \tag{15}$$

The algorithm iterates between the E-step and the M-step as discribed above until convergence. Cluster membership will then be assigned for each sample i based on the posterior mean of the latent variable $\boldsymbol{z}_i$ by a K-means approach. In the next section, we will discuss parameter tuning.

4.1. *Choice of Tuning Parameters.* We use a resampling-based criterion for selecting the penalty parameters and the number of clusters. The procedure entails repeatedly partitioning the data set into a learning and a test set. In each iteration, iCluster (for a given K and tuning parameter values) will be applied to the learning set to obtain a classifier and subsequently predict the cluster membership for the test set samples. Denote this as partition C_1. In parallel, the procedure applies an independent iCluster to the test set to obtain a second partition C_2. Under the true model, the predicted C_1 and the "truth" C_2 would have good agreement by measures such as the adjusted Rand index. We therefore define a reproducibility index (RI) as the median adjusted Rand index across all repetitions. Values of RI close to 1 indicate perfect cluster reproducibility and values of RI close to 0 indicate poor cluster reproducibility. In this framework, the concept of bias, variance, and prediction error that typically applies to classification analysis where the true cluster labels are known now becomes relevant for clustering. The idea is similar to the "Clest" method proposed by Dudoit and Fridlyand (2002), the prediction strength measure proposed by Tibshirani and Walther (2005), and the in-group proportion (IGP) proposed by Kapp and Tibshirani (2007).

4.2. *Uniform Sampling.* In the multidimensional space, an exhaustive grid search for the optimal combination of the penalty parameters and the number of clusters that maximizes cluster reproducibility is inefficient and computationally prohibitive. To improve efficiency, we use the uniform design approach of Fang and Wang (1994) to generate good lattice points from the parameter space, a similar strategy adopted by Wang *et al.* (2008). Let D be the search region. Using the concept of discrepancy that measures uniformity on $D \subset R^d$ with arbitrary dimension d, which is basically the Kolmogorov statistic for a uniform distribution on D, Fang and Wang (1994) point out that the discrepancy of the good lattice point set from a uniform design converges to zero with a rate of $O(n^{-1}(\log n)^d)$, here n (a prime number) denotes the number of generated points on D. They also point out that the sequence of equi-lattice points on D has a rate of $O(n^{-1/d})$ and the sequence of uniformly distributed random numbers on D has a rate of $O(n^{-1/2}(\log \log n)^{1/2})$. Thus the uniform design has an optimal rate for $d \geq 2$.

5. Application to Integrate Epigenomic and Transcriptomic Profiling Data in Breast Cancer. DNA methylation, an epigenetic modification which occurs most frequently in CpG sites, can result in heritable modification of gene expression in the absence of DNA sequence changes. It is an important epigenetic mechanism of transcriptional control. About 1% of human DNA consists of short areas containing CpG sites, and about half of all genes have a CpG island in their promoter region (Bird, 2002; Herman and Baylin, 2003; Egger *et al.*, 2004). Human cancer genomes are characterized by widespread aberrations in DNA methylation patterns. The challenge is to identify "driver" methylation changes that are essential to tumor initiation, progression or metastasis and distinguish these from methylation changes that are biologically neutral. Hypermethylation leads to epigenetic silencing (inactivation) of gene expression and has been associated with the inactivation of tumor suppressor genes. Therefore joint analysis of methylation and gene expression may provide important clue to driver gene identification and the cancer subtypes they define.

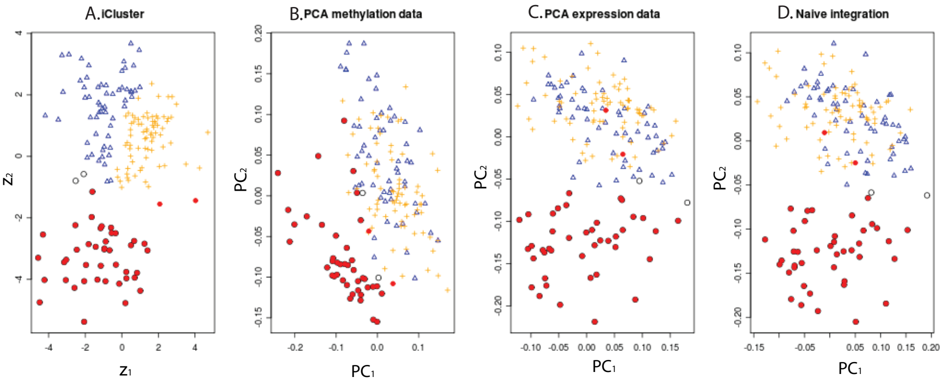

FIG 4. *Separation of the data points by A. iCluster latent variables in the integrated data space, B. the first two principal components in the methylation data alone, C. the first two principal components in the expression data alone, and D. a naive integration by data stacking followed by PCA analysis. Red dots indicate samples belonging to cluster 1, blue open triangles indicate samples belonging to cluster 2, and orange pluses indicate samples belonging to cluster 3.*

Holm *et al.* (2010) profiled methylation changes in 189 breast cancer samples using Illumina methylation array for 1,452 CpG sites (corresponding to 803 cancer-related genes) and performed hierarchical clustering on the methylation data alone. The authors then correlated the methylation status with gene expression level for 511 oligonucleotide probes for genes with CpG sites on the methylation assays in the same sample set. Here we applied iCluster to integrate the two data types. We obtained a 3-cluster (Figure 4A) solution that gives a reproducibility index of 0.70 with the combination of Lasso and Elasticnet penalty applied on methylation and gene expression data respectively (Table 1). Figure 4B and 4C indicate that a principal component analysis (PCA) in each data type alone can only separate one out of the three clusters. A naive integration by simply stacking the two data types (Figure 4D) also fails to separate the three clusters. Figure 5 reveals closely coordinated methylation and gene expression changes in tumors belonging to cluster 1 that is characterized by possible “driver” genes that are altered at both the epigenetic and transcriptomic level.

TABLE 1
Cluster reproducibility and number of genomic feature selected. K: the number of clusters. RI: reproducibility index.

	iCluster(Lasso, Lasso)			iCluster(Lasso, Elasticnet)		
K	RI	Selected methylation features	Selected expression features	RI	Selected methylation features	Selected expression features
2	0.68	138	151	0.70	183	353
3	0.46	150	204	0.70	273	182
4	0.42	183	398	0.48	273	182
5	0.42	205	454	0.47	282	223

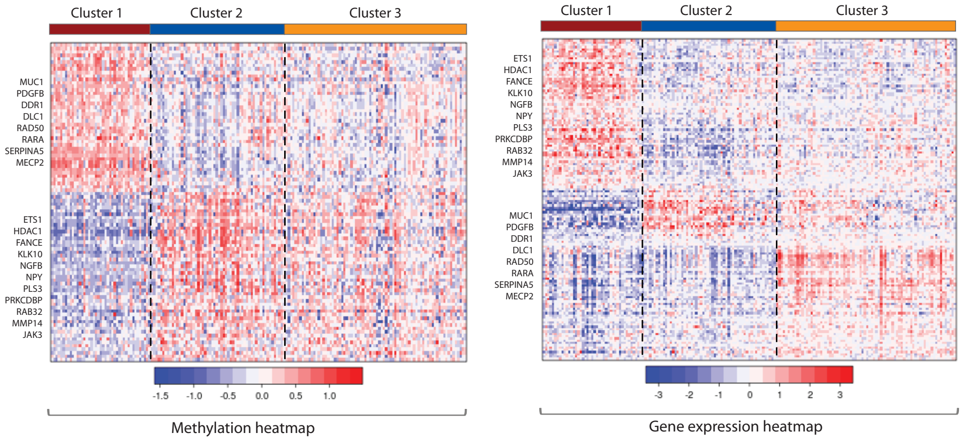

FIG 5. *Integrative clustering of the Holm study DNA methylation and gene expression data revealed three clusters with a cross-validated reproducibility of 0.7. Selected genes with negatively correlated methylation and expression changes are indicated to the left of the heatmap.*

6. Simulation.

In this section, we present results from two simulation studies. In the first simulation setup, we simulate one length-n latent factor $\boldsymbol{z} \sim N(0,1)$ where $n = 100$. Subject $i, i = 1, \cdots, n$ belongs to cluster 1 if $z_i > 0$ and cluster 2 otherwise. The coefficient vector $(\boldsymbol{w}_1, \boldsymbol{w}_2)$ is both of length $p_1 = p_2 = 200$ for simplicity (p_1 and p_2 can differ), with $w_{j1} = w_{j2} = 3$ and for $j = 1 - 20$ and zero elsewhere. Next we obtain the data matrices $(\boldsymbol{X}_1, \boldsymbol{X}_2)$ with each element generated according to equation (1).

Table 2

Clustering performance summarized over 50 simulated data sets under setup 1 (K=2). Separate K-means has two numbers associated with each criterion because of separate analysis of the paired data 1 and data 2. Number in parentheses is the standard deviation over 50 simulations.

Method	**Frequency of choosing the correct K**	**Cross-validation error rate**	**Cluster Reproducibility**
Separate K-means	58	0.08 (0.04)	0.67 (0.17)
	62	0.08 (0.04)	0.70 (0.19)
Standard iCluster	76	0.05 (0.03)	0.71 (0.16)
Lasso iCluster	**90**	**0.04 (0.02)**	**0.81 (0.08)**
Enet iCluster	**94**	**0.03 (0.02)**	**0.85 (0.07)**
Fused Lasso iCluster	**94**	**0.03 (0.02)**	**0.83 (0.08)**

Table 3

Feature selection performance summarized over 50 simulated data sets.

	Data 1		Data 2	
Method	**True positives**	**False positives**	**True positives**	**False positives**
Separate Kmeans	20(0)	10.2 (3.0)	20 (0)	10.1 (3.5)
Standard iCluster	20 (0)	8.0 (2.9)	20 (0)	8.6 (2.8)
Lasso iCluster	20 (0)	**0.07 (0.3)**	20 (0)	**0.07 (0.3)**
Enet iCluster	20 (0)	**0.1 (0.3)**	20 (0)	**0.02 (0.1)**
Fused Lasso iCluster	20 (0)	**0 (0)**	20 (0)	**0 (0)**

Table 2 summarizes the performances of each method in terms of the ability to choose the correct number of clusters, cross-validated error rates, cluster reproducibility. Table 3 summarizes the corresponding feature selection performance. In Table 2, the separate K-means clustering performs worst in terms of the ability to choose the correct number of clusters, cluster reproducibility, and the cross-validation error rates (with respect to the true simulated cluster membership). By contrast, sparse integrative clustering improves the performance by a large margin (bold numbers).

Table 3 summarizes the associated feature selection performance. We compared the traditional method of selecting features after clustering is done in the separate K-means approach as well as in the standard (nonsparse) iCluster where post-hoc univariate selection (based on t-test p-value cutoff) is applied. It is clear the difficulty with post-hoc univariate selection is false selection of noisy features in high dimensional data. In this simple simulation scenario, the Lasso iCluster method outperforms post-hoc univariate feature selection by keeping the number of falsely selected noisy features close to 0.

In the second simulation, we vary the setup as follows. We simulate 150 subjects belonging to three clusters ($K = 3$). Subject $i = 1 - 50$ belong to cluster 1, subject $i = 51 - 100$ belong to cluster 2, and subject $i = 101 - 150$ belong to cluster 3. A total of $m = 2$ data types ($\boldsymbol{X}_1, \boldsymbol{X}_2$) are simulated each has $p_1 = p_2 = 500$ features. Each data type alone only define two clusters out of the three. In data set 1, $x_{ij1} \sim N(2, 1)$ for $i = 1 - 50$ and $j = 1 - 10$, $x_{ij1} \sim N(1.5, 1)$ for $i = 51 - 100$ and $j = 491 - 500$, and $x_{ij1} \sim N(0, 1)$ for the rest. In data set 2, $x_{ij2} = 0.5 * x_{ij1} + e$ where $e \sim N(0, 1)$ for $i = 1 - 50$ and $j = 1 - 10$, $x_{ij2} \sim N(2, 1)$ for $i = 101 - 150$ and $j = 491 - 500$, and $x_{ij2} \sim N(0, 1)$ for the rest. The first 10 features are correlated between the two data types.

In Table 4, separate clustering fails miserably in terms of choosing the correct K, the error rate and cluster reproducibility due to the heterogeneity of the data types. The number of false positive features in Table 5 under the naive methods are overwhelmingly high. By contrast, the sparse iCluster methods show superior performance in correct identification of the true clusters and the true and false positive rates in feature selection.

TABLE 4
Clustering performance summarized over 50 simulated data sets under setup 2 (K=3).

Method	**Frequency of choosing the correct K**	**Cross-validation error rate**	**Cluster Reproducibility**
Separate K-means	2	0.33 (0.001)	0.54 (0.07)
	0	0.33 (0.002)	0.47 (0.04)
Standard iCluster	100	0.0007 (0.002)	0.98 (0.02)
Lasso iCluster	**100**	**0.0003 (0.001)**	**0.98 (0.01)**
Enet iCluster	**100**	**0.0003 (0.001)**	**0.97 (0.02)**
Fused Lasso iCluster	**100**	**0.0000 (0.000)**	**0.94 (0.05)**

TABLE 5
Feature selection performance summarized over 50 simulated data sets.

	Data 1		Data 2	
Method	**True positives**	**False positives**	**True positives**	**False positives**
Separate Kmeans	20 (0)	46.6 (7.5)	17.9 (1.7)	61.9 (6.7)
Standard iCluster	20 (0)	26.2 (7.6)	19.9 (0.3)	27 (6.5)
Lasso iCluster	20 (0)	**1.5 (1.4)**	19.9 (0.2)	**1.9 (1.5)**
Enet iCluster	20 (0)	**0.5 (0.6)**	19.8 (0.5)	**0.7 (1.0)**
Fused Lasso iCluster	20 (0)	**0 (0)**	20 (0)	**0 (0)**

6.1. *Implementation and running time.* The core iCluster EM iterations are implemented in C. Table 3 shows some typical computation times for problems of various dimensions on a 3.2 GHz Xeon Linux computer.

TABLE 6

Computing time (in seconds) for typical runs of sparse iCluster under various dimension.

		Time (in seconds)		
p	N	Lasso iCluster	Elasticnet iCluster	Fused Lasso iCluster
200	100	0.10	0.11	0.37
500	100	0.50	0.36	3.56
1000	100	1.40	1.45	25.05
2000	100	6.49	5.90	76.40
5000	100	18.93	18.94	33 (min)

7. Discussion.

We have formulated a joint latent variable model for integrating multiple genomic data sources. The latent variables are interpreted as a set of distinct underlying cancer driving factors that collectively explain the molecular phenotype manifested in the vast landscape of alterations in the cancer genome, epigenome, transcriptome. This is a novel concept for integrative class discovery we used to generate a single integrated cluster assignment based on simultaneous inference from multiple data types measured in the same set of patient samples. We derived a unified algorithm for a range of regularized minimization problems. The implementation scales well for increasing data dimension.

A natural extension of our current method would be to allow proper modeling of mixed data types when both discrete (mutation, SNP, sequencing count data) and continuous data are present. We are currently working on the generalized criteria to include mixtures of continuous and discrete data types appealing to properties of distributions in the exponential family.

References.

Alizadeh, A. A., Eisen, M. B., Davis, E. E. et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403** 503–511.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes and Development* **16** 6-21.

Cerami, E., Demir, E., Schultz, N., Taylor, B. S. and Sander, C. (2010). Automated network analysis identifies core pathways in glioblastoma. *PLoS one* **5** e8918.

Chin, L. and Gray, J. (2008). Translating insights from the cancer genome into clinical practice. *Nature* **452** 553-563.

Chitale, D., Gong, Y., Taylor, B. S., Broderick, S., Brennan, C., Somwar, R., Golas, B., Wang, L., Motoi, N., Szoke, J., Reinersman, J. M., Major, J., Sander, C., Seshan, V. E., Zakowski, M. F., Rusch, V., Pao, W., Gerald, W. and Ladanyi, M. (2009). An integrated genomic analysis of lung cancer reveals loss of DUSP4 in EGFR-mutant tumors. *Nature* **28** 2773-83.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **39** 1-38.

Dudoit, S. and Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology* **3** 1-21.

Egger, G., Liang, G., Aparicio, A. and Jones, P. A. (2004). Epigenetics in human disease and prospects for epigenetic therapy. *Nature* **429** 457-463.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalised likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348-1360.

Fang, K. T. and Wang, Y. (1994). *Number theoretic methods in statistics.* Chapman abd Hall, London, UK.

George, E. and Foster, D. (2000). Calibration and Empirical Bayes Variable Selection. *Biometrika* **87** 731-747.

George, E. and McCulloch, R. E. (1993). Variable Selection Via Gibbs Sampling. *Journal of the American Statistical Association* **88** 881-889.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Learning: Data Mining, Inference, and Prediction.* Springer, New York, NY.

Herman, J. G. and Baylin, S. B. (2003). Gene silencing in cancer in association with promoter hypermethylation. *New England Journal of Medicine* **349** 2042-2054.

Holm, K., Hegardt, C., Staaf, J. et al. (2010). Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns. *Breast Cancer Research* **12** R36.

Kapp, A. V. and Tibshirani, R. (2007). Are clusters found in one dataset present in another dataset? *Biostatistics* **8** 9-31.

Le Cao, K. A., Martin, P. G. and Robert-Granie, P.C. abd Besse (2009). Sparse canonical methods for biological data integration: application to a cross-platform study. *BMC Bioinformatics* **26** 34.

Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research* **8** 1145-64.

Park, T. and Casella, G. (2008). Calibration and Empirical Bayes Variable Selection. *Journal of the American Statistical Association* **103** 681-686.

Parkhomenko, E., Tritchler, D. and Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology* **8** 1-34.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D. Y., Pollack, J. R. and Wang, P. (2010). Regularized Multivariate Regression for Identifying Master Predictors with Application to Integrative Genomics Study of Breast Cancer. *Annals of Applied Statistics* **4** 53-77.

Perou, C. M., Jeffrey, S. S., van de Rijn, M. et al. (1999). Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proceedings of the National Academy of Sciences* **96** 9212-9217.

Pollack, J. R., Sørlie, T., Perou, C. M. et al. (2002). 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* **99** 12963–12968.

Sharpless, N. E. (2005). INK4a/ARF: a multifunctional tumor suppressor locus. *Mutat. Res.* **576** 99-102.

Shen, R., Olshen, A. B. and Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics* **25** 2906–2912.

Simon, R. (2010). Translational research in oncology: key bottlenecks and new paradigms. *Expert Reviews Molecular Medicine* **12**.

Soneson, C., Lilljebjrn, H., Fioretos, T. and Fontes, M. (2010). Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. *BMC Bioinformatics* **11** 191.

Sorlie, T., Perou, C. M., Tibshirani, R. et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences* **98** 10869–10874.

TCGA, (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455** 1061–1068.

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

Tibshirani, R. and Walther, G. (2005). Cluster Validation by Prediction Strength. *Journal of Computational & Graphical Statistics* **14** 511–528.

Tibshirani, R. and Wang, P. (2008). Spatial smoothing and hot spot detection for CGH data using the fused lasso. *Biostatistics* **1** 18–29.

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

Van Wieringen, W. N. and Van de Wiel, M. A. (2009). Nonparametric testing for DNA copy number induced differential mRNA gene expression. *Biometrics* **65** 19-29.

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, J.C. Zhu, Haussler, D. and M., S. J. (2010). Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26** 237-45.

Waaijenborg, S., Verselewel de Witt Hamer, P. C. and Zwinderman, A. H. (2008). Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis. *Statistical Applications in Genetics and Molecular Biology* **7** Article 3.

Wang, S. and Zhu, J. (2008). Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics* **64** 440–448.

Wang, S., Nan, B., Zhu, J. and Beer, D. (2008). Doubly Penalized Buckley-James Method for Survival Data with High-Dimensional Covariates. *Biometrics* **64** 132-140.

Witten, D. M. and Tibshirani, R. (2009). Extensions of sparse canonical correlation analysis, with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology* **8** Article 28.

Witten, D. M., Tibshirani, R. and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics* **10** 515-534.

Witten, D. M. and Tibshirani, R. (2010). A Framework for Feature Selection in Clustering. *Journal of the American Statistical Association* **105** 713-726.

Xie, B. and Pan, W. (2008). Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables. *Electronic Journal of Statistics* **2** 168-212.

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

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38** 894942.

Zhao, P., Rocha, G. and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics* **37** 3468-3497.

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

Department of Epidemiology and Biostatistics
Memorial Sloan-Kettering Cancer Center
New York, NY 10065
E-mail: shenr@mskcc.org

Department of Biostatistics and Medical Informatics
Department of Statistics
University of Wisconsin, Madison
Madison, WI 53792-4675
E-mail: swang@biostat.wisc.edu

Department of Epidemiology and Biostatistics
Memorial Sloan-Kettering Cancer Center
New York, NY 10065
E-mail: moq@mskcc.org