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
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, Verrelle de Witt Hamer and Zwinderman, 2008; Parkhomenko, Tritechler 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...
intrinsic/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.
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 \textit{HER2} tumor subtype (samples labeled in green), showing concordant amplification in the DNA and overexpression in mRNA at the \textit{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 \textit{HER2}-subtype specific genes (including \textit{HER2}, \textit{GRB7}, \textit{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 \textit{HER2} breast tumor subtype was revealed through the identification of the driving factor: \textit{HER2} activation. In general, if we have knowledge on the driving factors (such as \textit{HER2}) in a certain cancer and observe their values (degree of \textit{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 \( \mathbf{x}_{it} = (x_{i1t}, \cdots, x_{ipt})' \) denote a \( p_t \)-dimensional data vector. Each element \( x_{ijt}, j = 1, \cdots, p_t \), represents the observation associated with the \( j \)th genomic feature in data type \( t \) measured in tumor \( i \).

In matrix form, \( \mathbf{X}_t = (\mathbf{x}_{1t}, \cdots, \mathbf{x}_{nt}) \) is the \( t \)th 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 \( i \)th tumor. The reason we assume the length of \( \mathbf{z}_i \) 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 \( \mathbf{z}_i \sim MVN(0, \mathbf{I}) \). 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 \( \mathbf{w}_{jt} \) and \( \mathbf{z}_i \) 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

\[
(1) \quad x_{ijt} = \mathbf{w}_{jt}' \mathbf{z}_i + \epsilon_{ijt}, \quad i = 1, \cdots, n; \quad j = 1, \cdots, p_t; \quad t = 1, \cdots, m,
\]

where \( \mathbf{w}_{jt} \) is the length-(\( K - 1 \)) coefficient vector associated with the \( j \)th feature in the \( t \)th data type, and \( \mathbf{W}_t = (\mathbf{w}_{1t}, \cdots, \mathbf{w}_{pt})' \) is the corresponding coefficient matrix of dimension \( p_t \times (K-1) \). Finally, \( \epsilon_{it} \) denotes the error term with mean zero and covariance matrix \( \Psi_t = \text{diag}(\sigma_{1t}^2, \cdots, \sigma_{pt}^2) \), represent-
ing the residual variance. Marginally, \( x_{jt} \sim N(0, \Sigma_t) \) with \( \Sigma_t = W_t W_t^	op + \Psi_t \).

![Diagram](image)

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

The sparsity of \( W \) directly impacts the interpretability of the latent factors. If \( w_{kjt} = 0 \), then the \( k \)th latent variable has no effect on the \( j \)th feature in the \( t \)th 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 \( \hat{w}_{jt} \), we consider the following regularized log-likelihood:

\[
\ell_{c;p}(\{W_t\}_{t=1}^m, \{\Psi_t\}_{t=1}^m) = \ell_c - \sum_{t=1}^m J_{\lambda t}(W_t),
\]

where \( \ell_c \) is the complete-data log-likelihood function of the following form

\[
\ell_c \propto -\frac{n}{2} \sum_{t=1}^m \log |\Psi_t| - \frac{1}{2} \sum_{t=1}^m \sum_{i=1}^n (x_{it} - W_t z_i)\Psi_t^{-1}(x_{it} - W_t z_i) - \frac{1}{2} \sum_{i=1}^n z_i' z_i,
\]

In (2), \( \ell_c \) controls the fitness of the model, and \( J_{\lambda t}(W_t) \) is a penalty function which controls the complexity of the model; and \( \lambda_t \)'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 \( \ell_1 \)-penalty (Tibshirani, 1996) that takes the form

\[
J_{\lambda t}(W_t) = \lambda_t \sum_{k=1}^{K-1} \sum_{j=1}^m |w_{kjt}|.
\]

The \( \ell_1 \)-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 \( \ell_1 \)-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 $\lambda_t$. If the entire row-vector $w_{tj}$ 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

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

where $\lambda_{1t}$ and $\lambda_{2t}$ 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

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

where $\lambda_{1t}$ and $\lambda_{2t}$ 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 $(w_1, w_2)$ 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 z_i. Specifically, in the E-step, we compute the quantity

\[ E\left[ \ell_{c,p} \mid \{X_i\}_{i=1}^m, \{W_t^{(q)}\}_{t=1}^m, \{\Psi_t^{(q)}\}_{t=1}^m \right]. \]
In the M-step, for \( t = 1, \ldots, m \), we obtain the penalized estimates by

\[
\hat{W}_t \leftarrow \arg\min_{W_t} \frac{1}{2} \sum_{i=1}^{n} E \left[ (x_{it} - W_t z_i)^\prime \Psi_t^{-1} (x_{it} - W_t z_i) \right] + J_t(W_t).
\]

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

\[
\sum_{i=1}^{n} x_{it} E[\zeta_i'|x_{it}, W_t^{(q)}, \Psi_t^{(q)}] - W_t \sum_{i=1}^{n} E[z_i z_i'|x_{it}, W_t^{(q)}, \Psi_t^{(q)}] - \Psi_t \frac{\partial}{\partial W_t} J(W_t) = 0.
\]

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

\[
C = \sum_{i=1}^{n} x_{it} E[\zeta_i'|x_{it}, W_t^{(q)}, \Psi_t^{(q)}], \quad \text{and} \quad Q = \sum_{i=1}^{n} E[z_i z_i'|x_{it}, W_t^{(q)}, \Psi_t^{(q)}].
\]

Then the above becomes

\[
\Psi \frac{\partial}{\partial W_t} J(W_t) + WQ = C.
\]

Due to the diagonal structure of \( \Psi \) and “non-coupling” structure of the penalty term, the estimation of \( w_j \)'s are separable. In other words, we can consider the following \( p \) estimating equations:

\[
\sigma_j^2 \frac{\partial}{\partial w_j} J(w_j) + w_j' Q = C_j, \quad j = 1, \ldots, p,
\]

where \( C_j \) is the \( j \)th row of \( C \). Or equivalently,

\[
\sigma_j^2 \frac{\partial}{\partial w_j} J(w_j) + Q' w_j = C_j', \quad j = 1, \ldots, p.
\]

However, the nondifferentiability of the objective function involving \( \ell_1 \) 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 \( \ell_1 \) term:

\[
(8) \quad J_{\lambda}(W_t)_{\text{lasso}} \approx \lambda \sum_{k=1}^{K-1} \sum_{j=1}^{p_t} \frac{w_{kjt}^2}{|\hat{w}_{kjt}^{(q)}|}.
\]
and obtain the solution for (7) under Lasso penalty by iteratively computing the following ridge type regression

\[
w_{jt} = \left( \frac{1}{n} \sum_{i=1}^{n} E[z_i z_i' | x_{it}, W_t^{(q)}, \Psi_t^{(q)}] + A_j \right)^{-1} \sum_{i=1}^{n} x_{itj} E[z_i' x_{it}, W_t^{(q)}, \Psi_t^{(q)}],
\]

where \( A_j = 2 \sigma_j^2 \lambda \text{diag}\{1/|\hat{w}_{1j}|, \ldots, 1/|\hat{w}_{(K-1)j}|\} \) for \( j = 1, \ldots, p_t \). Unlike the standard Ridge regression estimates typically require the inversion of a \( p \times p \) 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 \( \ell_1 \) term in the Elasticnet penalty:

\[
J_{\lambda_t}^{(W_t)} \approx \lambda_{1t} \sum_{k=1}^{K-1} \sum_{j=1}^{p_t} \frac{w^{2}_{kj}}{\hat{w}^{(q)}_{kj}} + \lambda_{2t} \sum_{k=1}^{K-1} \sum_{j=2}^{p_t} \frac{(w_{kj} - w_{k(j-1)t})^2}{\hat{w}^{(q)}_{kj} - \hat{w}^{(q)}_{k(j-1)t}}.
\]

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

\[
J_{\lambda_t}^{(w_t)} \approx \lambda_{1t} \sum_{k=1}^{K-1} \sum_{j=1}^{p_t} \frac{w^{2}_{kj}}{|\hat{w}^{(q)}_{kj}|} + \lambda_{2t} \sum_{k=1}^{K-1} \sum_{j=2}^{p_t} \frac{(w_{kj} - w_{k(j-1)t})^2}{|\hat{w}^{(q)}_{kj} - \hat{w}^{(q)}_{k(j-1)t}|}.
\]

\[
= \lambda_{1t} w_t' A w_t + \lambda_{2t} w_t' L w_t,
\]
where

\[
A = \text{diag}\left\{1/|\hat{w}_1^{(q)}|, \ldots, 1/|\hat{w}_s^{(q)}|\right\} \quad (s = p_t \cdot (K - 1)),
\]

\[
L = D - M,
\]

\[
M = \left\{ \begin{array}{l} 1/|\hat{w}_i^{(q)} - \hat{w}_j^{(q)}|, \quad |i - j| = K - 1 \\ 0, \quad \text{otherwise.} \end{array} \right. \quad (s \times s \text{ dimension}),
\]

\[
D = \text{diag}\{D_1, \ldots, D_s\} \quad \text{with} \quad D_j \text{ be the summation of the } j \text{ th row of } M.
\]

The corresponding representation of estimating equation is

\[
\frac{\partial}{\partial w} J(w) + \tilde{Q}w = \tilde{C},
\]

where

\[
\tilde{Q} = \begin{pmatrix} \sigma_1^{-2}Q & \cdots & 0 \\
0 & \ddots & 0 \\
0 & 0 & \sigma_p^{-2}Q \end{pmatrix}, \quad \tilde{C} = \begin{pmatrix} \sigma_1^{-2}C_1' \\
\vdots \\
\sigma_p^{-2}C_p' \end{pmatrix}.
\]

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

\[
w_t = \left( \tilde{Q} + 2\lambda_1 A + 2\lambda_2 L \right)^{-1} \tilde{C}.
\]

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 \( \Psi \) in the M-step as follows

\[
\hat{\Psi}_t = \frac{1}{n} \text{diag}(X_tX_t' - W_t^{(q)}E[Z|\{X_t\}_{t=1}^m, \{W_t^{(q)}\}_{t=1}^m, \{\Psi_t^{(q)}\}_{t=1}^m]X_t').
\]

The algorithm iterates between the E-step and the M-step as described above until convergence. Cluster membership will then be assigned for each sample \( i \) based on the posterior mean of the latent variable \( z_i \) 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 \( D \subset \mathbb{R}^d \) 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 character-
ized 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.

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.

<table>
<thead>
<tr>
<th>K</th>
<th>RI</th>
<th>Selected methylation features</th>
<th>Selected expression features</th>
<th>K</th>
<th>RI</th>
<th>Selected methylation features</th>
<th>Selected expression features</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.68</td>
<td>138</td>
<td>151</td>
<td>0.70</td>
<td>183</td>
<td>155</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0.46</td>
<td>150</td>
<td>204</td>
<td>0.70</td>
<td>273</td>
<td>182</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.42</td>
<td>183</td>
<td>398</td>
<td>0.48</td>
<td>273</td>
<td>182</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>0.42</td>
<td>205</td>
<td>454</td>
<td>0.47</td>
<td>282</td>
<td>223</td>
<td></td>
</tr>
</tbody>
</table>

![Methylation heatmap](image1)

![Gene expression heatmap](image2)

**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 $z \sim N(0, 1)$ where $n = 100$. Subject $i, i = 1, \ldots, n$ belongs to cluster 1 if $z_i > 0$ and cluster 2 otherwise. The coefficient vector $(w_1, w_2)$ is both of length $p_1 = p_2 = 200$ for simplicity ($p_1$ and $p_2$ can differ), with $w_{j_1} = w_{j_2} = 3$ and for $j = 1 - 20$ and zero elsewhere. Next we obtain the data matrices $(X_1, X_2)$ with each element generated according to equation (1).
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 \((X_1, X_2)\) 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 \((K=3)\).**

<table>
<thead>
<tr>
<th>Method</th>
<th>Frequency of choosing the correct (K)</th>
<th>Cross-validation error rate</th>
<th>Cluster Reproducibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Separate K-means</td>
<td>2</td>
<td>0.33 (0.001)</td>
<td>0.54 (0.07)</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.33 (0.002)</td>
<td>0.47 (0.04)</td>
</tr>
<tr>
<td>Standard iCluster</td>
<td>100</td>
<td>0.0007 (0.002)</td>
<td>0.98 (0.02)</td>
</tr>
<tr>
<td>Lasso iCluster</td>
<td>100</td>
<td>0.0003 (0.001)</td>
<td><strong>0.98 (0.01)</strong></td>
</tr>
<tr>
<td>Enet iCluster</td>
<td>100</td>
<td>0.0003 (0.001)</td>
<td>0.97 (0.02)</td>
</tr>
<tr>
<td>Fused Lasso iCluster</td>
<td>100</td>
<td>0.0000 (0.000)</td>
<td><strong>0.94 (0.05)</strong></td>
</tr>
</tbody>
</table>

### Table 5

**Feature selection performance summarized over 50 simulated data sets.**

<table>
<thead>
<tr>
<th>Method</th>
<th>Data 1</th>
<th>Data 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>True positives</td>
<td>False positives</td>
</tr>
<tr>
<td>Separate K-means</td>
<td>20 (0)</td>
<td>46.6 (7.5)</td>
</tr>
<tr>
<td>Standard iCluster</td>
<td>20 (0)</td>
<td>26.2 (7.6)</td>
</tr>
<tr>
<td>Lasso iCluster</td>
<td>20 (0)</td>
<td><strong>1.5 (1.4)</strong></td>
</tr>
<tr>
<td>Enet iCluster</td>
<td>20 (0)</td>
<td><strong>0.5 (0.6)</strong></td>
</tr>
<tr>
<td>Fused Lasso iCluster</td>
<td>20 (0)</td>
<td>0 (0)</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>p</th>
<th>N</th>
<th>Time (in seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Lasso iCluster</td>
</tr>
<tr>
<td>200</td>
<td>100</td>
<td>0.10</td>
</tr>
<tr>
<td>500</td>
<td>100</td>
<td>0.50</td>
</tr>
<tr>
<td>1000</td>
<td>100</td>
<td>1.40</td>
</tr>
<tr>
<td>2000</td>
<td>100</td>
<td>6.49</td>
</tr>
<tr>
<td>5000</td>
<td>100</td>
<td>18.93</td>
</tr>
</tbody>
</table>

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.


