Integrative clustering of high-dimensional data with joint and individual clusters

KRISTOFFER H. HELLTON *
Oslo Center for Biostatistics and Epidemiology, Department of Biostatistics, University of Oslo, N-0317 Oslo, Norway
Division of Medicine and Laboratory Sciences, Institute of Clinical Medicine, University of Oslo, N-1478 Lørenskog, Norway
k.h.hellton@medisin.uio.no

MAGNE THORESEN
Oslo Center for Biostatistics and Epidemiology, Department of Biostatistics, University of Oslo, N-0317 Oslo, Norway

SUMMARY
When measuring a range of genomic, epigenomic and transcriptomic variables for the same tissue sample, an integrative approach to analysis can strengthen inference and lead to new insights. This is also the case when clustering patient samples, and several integrative cluster procedures have been proposed. Common for these methodologies is the restriction to a joint cluster structure, equal in all data layers. We instead present a clustering extension of the JIVE algorithm (Lock and others, 2013), Joint and Individual Clustering (JIC), enabling the construction of both joint and data type-specific clusters simultaneously. The procedure builds on the connection between k-means clustering and principal component analysis, and hence, the number of clusters can be determined by the number of relevant principal components. The proposed procedure is compared to iCluster, a method restricted to only joint clusters, and simulations show that JIC is advantageous when both individual and joint clusters are present. The procedure is illustrated using gene expression and miRNA levels measured in breast cancer tissue from the The Cancer Genome Atlas (TCGA). The analysis suggests a division into three joint clusters common for both data types and two expression-specific clusters.

Key words: Breast cancer; Clustering; Integrative genomics; Latent variable estimation; Singular value decomposition.

1. INTRODUCTION
The rapid development in genomic technologies has enabled the measurement of a range of genomic, epigenomic and transcriptomic data layers or data types. This increases the need for

*To whom correspondence should be addressed.
integrative procedures that can handle several data types simultaneously. When studying diseases building on molecular processes, one needs to consider the interplay between the genomic layers to fully understand the phenotypic traits, and one should attempt to integrate different data types in a single joint analysis. As the information content is higher in an integrative framework compared to individual analyses, it is possible to gain statistical power to detect relevant signals. An integrative approach is particularly relevant for unsupervised clustering in genetically driven diseases such as cancer, where the aim of clustering is to discover novel disease subtypes to aid the understanding of survival and mortality risk differences.

The aim of integrative clustering is to identify a set of joint clusters for the considered data sources. Currently, there are several proposed methodologies spanning different directions; joint latent variable modeling, such as the iCluster method (Shen and others, 2009, 2013), non-negative matrix factorization (Zhang and others, 2012), Gaussian mixture-models (Kormaksson and others, 2012) and several Bayesian approaches (Kirk and others, 2012; Lock and Dunson, 2013; Ray and others, 2014). An overview of these different directions is given by Chalise and others (2014).

Common for these methodologies is the aim of finding a joint cluster structure, which is equal in all data layers. However, when trying to identify clusters in highly heterogeneous genomic data, some cluster structures are typically not shared between all the data layers. If there are clear clusters present in some of the data types, but not in others, these can confound or obscure the joint clusters shared by all data types. Data type-specific cluster structures could be caused by (uninteresting) technical or measurement-related issues, affecting only a single data type. But there could also exist disease-related patient clusters, that are independent of the joint clusters and still relevant for treatment and disease-understanding.

Currently, several approaches to decompose multisource data into common and data type-specific components are being developed (Lock and others, 2013; Schouteden and others, 2013; Löfstedt and others, 2013). Our aim is to take into account the presence of data type-specific clusters together with the joint clusters by utilizing such a decomposition technique. We present a clustering procedure based on the Joint and Individual Variance Explained (JIVE) algorithm (Lock and others, 2013), where common and data type-specific structures are assumed additive and orthogonal. In our extension, termed Joint and Individual Clustering (JIC), the joint and data type-specific cluster structures are estimated simultaneously. Our proposed procedure is connected to the iCluster methodology, and these methods are compared in different simulation settings. Lastly, JIC will be used to find joint and data type-specific clusters in breast cancer tissue samples.

2. INTEGRATIVE CLUSTERING

In integrative clustering, the aim is to cluster observations using several data types. Let $X_1, \ldots, X_M$ be $M$ different genome-scale data types (typically expression, copy number variation, methylation) or genome-related data types (such as miRNA, proteins, transcription factors) that are all measured on the same $n$ patients, indexed $j = 1, \ldots, n$. For $m = 1, \ldots, M$, each $X_m$ is a $p_m \times n$ data matrix with $p_m$ variables, indexed by $i = 1, \ldots, p_m$. The data types can be highly heterogeneous with respect to scale, unit or variation. The $M$ data matrices can be combined into a single concatenated matrix

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_M \end{bmatrix}.$$
of dimension $p \times n$ where $p = p_1 + \cdots + p_M$. A scaled version of the concatenated matrix can be constructed by first scaling each data matrix $X_m$ by the Frobenius norm, $\|X_m\|_F$, (Lock and others, 2013).

### 2.1 Clustering and dimension reduction

We extend the JIVE algorithm (Lock and others, 2013) to accommodate clustering by utilizing the connection between k-means clustering and principal component analysis (PCA). The basic k-means clustering procedure finds an optimal division of the observations into $K$ groups by minimizing the distance between each observation and the group centroid. To simplify the procedure of k-means, PCA has been used as an initial step to reduce the dimension of the data matrix. This two-step procedure clusters the principal component scores of the most informative components, but has been criticized in the statistics literature (Arabie and Hubert, 1996; Chang, 1983) due to the lack of a formal connection between the number of chosen components and number of clusters.

However, in the field of machine learning, Zha and others (2001) and Ding and He (2004) have shown that the principal components are in fact the continuous solution to the k-means optimization problem. The principal component scores will therefore correspond to a continuous version of the discrete cluster indicator vectors. Specifically, if the k-means clustering solution for $K$ clusters is denoted by a matrix of $K - 1$ indicator vectors $\tilde{Z} = [\tilde{z}_1, \ldots, \tilde{z}_{K-1}]^T$ and

$$\tilde{z}_k^T = n_k^{-1/2} [0, \ldots, 0, 1, \ldots, 1, 0, \ldots, 0],$$

where $n_k$ is the number of observations in each cluster, the $K - 1$ first principal component score vectors will be the continuous solution minimizing the k-means objective function (Ding and He, 2004, Theorem 3.1). Therefore, k-means clustering into $K$ groups can be solved in two steps: first find the $K - 1$ (standardized) principal component scores, and then reconstruct the discrete cluster assignments from the continuous scores, for instance with k-means clustering. For high-dimensional data, the initial dimension reduction step greatly increases the efficiency of the clustering step, as the dimension of the $n$ observations is reduced from $p$ to $K - 1$.

### 2.2 The iCluster method

The iCluster method (Shen and others, 2009, 2013) utilizes this connection, but incorporates Gaussian latent variable modeling to obtain an integrative procedure. In latent variable modeling, the data matrix $X_m$ is given as

$$X_m = W_m Z + \varepsilon_m, \quad \varepsilon_m \sim N(0, \Sigma),$$

where $W_m$ is a $p \times (K - 1)$ loading coefficient matrix, $Z$ is the latent variable matrix and $\varepsilon_m$ is a set of independently distributed errors. Tipping and Bishop (1999) connected Gaussian latent variable models and PCA by showing that, for homogeneous errors $\Sigma = \sigma^2 I_p$, the maximum likelihood estimates of the loading coefficient matrix yields a solution equivalent to PCA. The iCluster method assumes the clusters, represented by the latent variable matrix $Z$, to be common
for all data types
\[
X_1 = W_1 Z + \varepsilon_1,
\]
\[
\vdots
\]
\[
X_M = W_M Z + \varepsilon_M,
\]
and allows for heterogeneous noise terms, \( \varepsilon_m \sim N(0, \Sigma_m) \), \( \Sigma_m = \text{diag}(\sigma^2_1, \ldots, \sigma^2_{p_m}) \). The parameter estimates are obtained by maximum likelihood estimation using the EM algorithm, where sparse loading matrices can enforced by penalizing the data log-likelihood. After convergence of the EM algorithm, the rows of \( Z^T \) are clustered by the k-means algorithm to obtain the group membership of each observation, such that the latent variable, \( Z \), corresponds to a cluster indicator matrix shared between all data types. This approach to integration is also used in simultaneous component analysis (Deun and others, 2009, 2011).

2.3 Joint and Individual Clustering (JIC)

Our proposed procedure, Joint and Individual Clustering (JIC), utilizes the connection between k-means clustering and PCA within the JIVE algorithm (Lock and others, 2013), where PCA (or the singular value decomposition) is used to decompose the concatenated data matrix into orthogonal joint and individual matrices. In JIC, the resulting shared and the data type-specific principal component scores are both used to obtain joint and data type-specific clusters in a finale k-means step.

In the latent variable formulation, JIC assumes the data matrices to be given by common clusters \( Z \) and data type-specific clusters \( Z_1, \ldots, Z_M \)
\[
X_1 = W_1 Z + V_1 Z_1 + \varepsilon_1,
\]
\[
\vdots
\]
\[
X_M = W_M Z + V_M Z_M + \varepsilon_M,
\]
where \( \varepsilon_m \sim N(0, \sigma^2_m I), m = 1, \ldots, M \) and the \( p_m \times (K - 1) \) joint loading matrices form a \( p \times (K - 1) \) concatenated matrix \( W = [W_1, \ldots, W_M]^T \).

The JIVE algorithm constructs the same decomposition by minimizing the reconstruction error
\[
\sum_{m=1}^M \|X_m - W_m Z - V_m Z_m\|^2,
\]
while assuming each \( Z_m \) to be orthogonal to \( Z \), \( ZZ_m^T = 0_{(K-1)\times(K-1)} \) for \( m = 1, \ldots, M \) to ensure a unique solution (Lock and others, 2013, Supplementary material). When the ranks of \( WZ \) and \( V_m Z_m \), \( r \) and \( r_m \) for \( m = 1, \ldots, M \), are given, JIVE decomposes the concatenated matrix by iteratively estimating the joint structures with the individual structures fixed and vice versa, until a suitable convergence criterion is reached. The Joint and Individual Clustering is then given by the following two-step procedure:

1. Estimate the joint and individual matrices, \( WZ \) and \( V_m Z_m \) using the JIVE algorithm. Initialize \( X^{\text{JOINT}} = X \), repeat the following steps until convergence,
   - set \( WZ \) to be the \( r \) rank singular value decomposition of \( X^{\text{JOINT}} \),
Joint and individual clustering

• for \( m = 1, \ldots, M \), set \( X^{\text{INDIVID}}_m = X_m - W_m Z \), set \( V_m Z_m \) to be the \( r_m \) rank singular value decomposition of \( X^{\text{INDIVID}}_m (I - Z Z^T) \),

• form the concatenated matrix of \( X^{\text{JOINT}} = [X_1 - V_1 Z_1, \ldots, X_M - V_M^T Z_M]^T \).

After convergence, \( Z \) is given by the \( r \) first right singular vectors of \( X^{\text{JOINT}} \) and \( Z_m \) is given by the \( r_m \) first right singular vectors of \( X^{\text{INDIVID}}_m \) for \( m = 1, \ldots, M \).

2. Cluster the rows of \( Z^T \) into \( r + 1 \) groups and the rows of \( Z^T_m \) into \( r_m + 1 \) groups for \( m = 1, \ldots, M \) using k-means clustering.

Both JIC and iCluster estimate the continuous representations of the cluster assignment vectors, but they differ in terms of the assumed noise structure. iCluster allows the variance of the noise in each variable to be different, while the JIVE algorithm, and thereby JIC, is restricted to equal variances to ensure identifiability of the joint and individual components via the singular value decomposition.

2.4 Procedure for selection of the number of clusters

With the connection \( K = r + 1 \) (Ding and He, 2004), it is possible to utilize selection procedures for the number of clusters to select the rank of the latent structures. The configuration of the numbers of clusters \( K, K_1, \ldots, K_m \) gives the ranks \( r, r_1, \ldots, r_m \), used in the JIVE decomposition. A requirement is, however, that the selection procedure is able to assess whether no clusters are present, i.e. \( K = 1 \) and \( r = 0 \).

Generally, there is no optimal procedure for selecting the number of clusters, but the classical approach for k-means clustering (Kaufman, 2009) is to evaluate the cluster separation, for instance measured by the Calinski-Harabasz criterion, Dunn criterion or C-index. The cluster separation is calculated for a range of different values of \( K \) and the \( K \) with the best value is chosen. This approach could also be used over a grid of values for \( K \) and \( K_1, \ldots, K_m \), but the number of possible combinations grows exponentially in \( M \). If \( M = 2 \) and \( K, K_1 \) and \( K_2 \) range from 1 to 10, there are \( 10^3 \) different combinations to compare, and the extreme number of observations make an assessment difficult. In addition, the values of the separation criterion will depend on the data characteristics, thus there is no straightforward way of combining criterion values calculated for different data types to find an overall optimal solution. The approach did not yield satisfactory behavior in neither simulations nor in real data.

An alternative approach is to use a prediction strength criterion (Tibshirani and Walther, 2005; Shen and others, 2013) based on cluster reproducibility. Shen and others (2013) evaluated the predictive power for different \( K \) by randomly splitting the data into discovery and validation sets and measuring the similarity between the prediction and validation clusterings. Analogous to cross-validation, the approach assesses whether specific clusters can be reproduced in subsets of the data. As the criterion measures differences on cluster level (e.g. the proportion of correct predictions), the criterion values are less sensitive to the data distribution and are easier to compare between different types of data. However, as the procedure chooses the number of clusters to give good cluster reproducibility, the cluster separation is not necessarily taken into account. In the high-dimensional setting, the predicted scores of the components with the largest eigenvalues tend to be very stable (Hellton and Thoresen, 2014) regardless of sub-sampling. This results in a perfect reproducibility for small \( K \), making it difficult to select an optimal \( K \).

Instead we utilize the result of Ding and He (2004): when optimizing the k-means objective function for \( K \) clusters, the continuous solution is given by the \( K - 1 \) first principal components. We will therefore recover the number of clusters by identifying the principal components where
new clusters are present. Our approach is to check if a new cluster is separated out in each added component, until we find the first component where only a single cluster is present:

1. For the $i$th component, check if there are more than one cluster present, by some procedure.

2. If so, proceed to the next component. If not, stop and set the number of clusters $K$ to be the current component number and the rank to be $r = K - 1$.

When using PCA to cluster gene expression, data clusters tend to be less well-separated (Yeung and Ruzzo, 2001) and instead resemble a continuum. One approach that can handle this situation is to evaluate the normality of the scores, as done in the G-means procedure by Hamarly and Elkan (2003). Under the assumption of normally distributed noise, the component scores will be normally distributed only if there is no cluster structure. The normality can be assessed by normality tests, but we prefer evaluation of density and quantile-quantile (qq) plots. If the distribution of the component scores deviates from normality, the component does not represent noise and clusters should be present.

To ensure identifiability, the JIVE procedure imposes the restrictions that the joint and individual matrices, $WZ$ and $V_mZ_m$, are orthogonal, and that the intersection of the row spaces of $V_1^tZ_1$ to $V_MZ_M$ is empty (Lock and others, 2013, Supplementary material), such that the rank of $WZ$ and the concatenated $V_mZ_m$ in the population model is given as

$$\text{rank} \begin{bmatrix} W_1Z + V_1Z_1 \\ \vdots \\ W_MZ + V_MZ_M \end{bmatrix} = r + r_1 + \cdots + r_M, (2.1)$$

while the rank of the sub-matrices $W_mZ$ and $V_mZ_m$ will be

$$\text{rank}(W_mZ + V_mZ_m) = r + r_m, \quad m = 1, \ldots, M. (2.2)$$

We assume that k-means clustering is an appropriate method for recovering the true clusters, such that a component representing noise cannot be larger than a component representing a cluster. Then the rank in Equation (2.1), denoted $E$, can be found by the number of clusters in the concatenated matrix, $X$. Similarly, the ranks in Equation (2.2), denoted $E_m$, can be found by the number of clusters in the original data $X_m$. With an estimate of $E$ and $E_1, \ldots, E_M$, Equations (2.1) and (2.2) determine a system of equations, where $r$ and $r_1, \ldots, r_m$ are given as

$$r = \frac{E_1 + \cdots + E_M - E}{M - 1}, \quad r_m = E_m - r, \quad m = 1, \ldots, M. (2.3)$$

To find the numbers of clusters, $K$ and $K_1, \ldots, K_M$, we use the following procedure:

1. Assess the number of relevant components $E$ in $X$: check the normality of the $i$th component score vector of $X$ for increasing $i$, until the last non-normally distributed component is found and set $E$ to the component number.

2. Assess the number of relevant subspaces $E_m$ in $X_m$: For each $m = 1, \ldots, M$, check the normality of the $i$th component score vector of $X_m$ for increasing $i$ up to $E$, until the last non-normally distributed component is found, and set $E_m$ to the component number.

3. Based on $E, E_1, \ldots, E_M$, calculate the ranks $r, r_1, \ldots, r_m$ by Equation (2.3) and set the numbers of clusters to be $K = r + 1$ and $K_m = r_m + 1$ for $m = 1, \ldots, M$. 
3. Simulations

We compare JIC to the iCluster procedure in two different simulation settings; only joint clusters and both joint and data type-specific clusters. In both settings, three different data types are integrated, \( M = 3 \), and the number of clusters is first assumed known, then estimated by the procedure described in Section 2.4.

3.1 Setting I: Joint cluster structure

First, we simulate 5 joint clusters present in all three data sets. For \( n = 150 \), the observations \( j = 1, \ldots, 30 \) belong to the first cluster, \( j = 31, \ldots, 60 \) belong to the second cluster and so on, giving 30 observations in each cluster. The joint latent variable, \( Z^T_J \), is an \( n \times 4 \) matrix with the indicator vectors as columns

\[
Z^T_J = \begin{bmatrix}
1 & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots \\
0 & 1 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots 
\end{bmatrix}.
\]

Each row contains a single ‘1’ indicating the assignment of the observation to the cluster corresponding to the column number. The last cluster is, however, specified by only zeros. The set of loading matrices \( W_1, W_2 \) and \( W_3 \) are of the same dimension \( 200 \times 4 \) (\( p_1 = p_2 = p_3 = 200 \)). We generate the loadings according to a standard normal distribution and normalize the matrices, such that \( W^T_m W_m = I \) for \( m = 1, 2, 3 \). The three data sets are generated by

\[
X_1 = cW_1Z_J + \varepsilon_1,
\]
\[
X_2 = cW_2Z_J + \varepsilon_2,
\]
\[
X_3 = cW_3Z_J + \varepsilon_3,
\]

with standard normally distributed errors, \( \varepsilon_m \sim N(0, I) \), and \( c = 80 \).

In the simulations, we first assume \( K = 5 \) known and compare the estimated cluster assignments to the true clusters in terms of the precision. Secondly, we assume the number of clusters unknown and estimate \( K \) by the procedure in Section 2.4. Table 1 shows the precision of JIC compared to the iCluster procedure in the two simulation settings. In Setting I, we see that iCluster and JIC perform equally well in finding the joint clusters due to the homogeneous noise structure. In the case of an unknown number of clusters, the selection procedure correctly estimates \( K \) to be 5 in 97% of the simulated cases, as seen in Table 1. The fact that there are no individual clusters present in any of the three data sets is correctly estimated in 96%, 95% and 98% of the simulated data sets, respectively.

3.2 Setting II: Joint and individual clusters

In Setting II, two data type-specific clusters are added to each of the three data sets. The observations are randomly assigned to one of two clusters, such that the data type-specific latent variables \( Z_1, Z_2 \) and \( Z_3 \) are vectors with random ones and zeros. For the loading matrices \( V_1, V_2 \) and \( V_3 \) of dimension \( 200 \times 1 \), the loadings are randomly generated according to a standard normal distribution and normalized, such that \( V^T_m V_m = 1 \) for \( m = 1, 2, 3 \).

To obtain an identifiable decomposition, each \( Z_m \) is made orthogonal to the columns of \( Z_J \).
K. H. Hellton & M. Thoresen

The three data sets are generated by the model

\[ X_1 = cW_1Z_J + c_1V_1Z_1 + \varepsilon_1, \]
\[ X_2 = cW_2Z_J + c_2V_2Z_2 + \varepsilon_2, \]
\[ X_3 = cW_3Z_J + c_3V_3Z_3 + \varepsilon_3, \]

with standard normally distributed noise, \( \varepsilon_m \sim N(0, I) \) and four different set-ups for the relative signal strengths of the joint clusters, \( c \), and the individual clusters, \( c_1, c_2, c_3 \):

1. \( c = 80, \ c_1 = c_2 = c_3 = 32 \)
2. \( c = 80, \ c_1 = c_2 = 32, \ c_3 = 16 \)
3. \( c = 80, \ c_1 = 32, \ c_2 = c_3 = 16 \)
4. \( c = 80, \ c_1 = c_2 = c_3 = 16 \)

In the first set-up, all three of the data-specific clusters obscure the joint clusters and in the second and third set-up, the first two and only the first data-specific clusters obscure the joint clusters, respectively. In the fourth set-up, none of data-specific clusters obscure the joint clusters as \( c_1, c_2, c_3 \) are too small compared to \( c \).

First, the correct numbers of clusters, \( K = 5 \) and \( K_1 = K_2 = K_3 = 2 \), are assumed known and the joint and individual clustering are compared to the true cluster memberships with the precisions shown in Table 1. As the iCluster procedure only finds joint clusters, the precisions for the individual clusters are not shown. It is seen that as more of the individual clusters obscure the joint clusters, JIC shows increasingly better performance compared to iCluster. Furthermore, JIC achieves a high precision for both the joint and individual clusters, regardless of their relative signal strength. Then, in the case of unknown \( K \) and \( K_1, K_2, K_3 \), it is seen from Table 1 that the numbers of clusters are most often correctly estimated. Only in the fourth set-up, where iCluster and JIC perform equally well, are the signal strength of the individual clusters too small, compared to the joint clusters, to properly estimate the numbers of clusters.

4. Example: expression and miRNA data

To illustrate our clustering procedure, we use data from The Cancer Genome Atlas (TCGA), specifically gene expression and miRNA measured in breast cancer tissue (Cancer Genome Atlas Network, 2012). We restrict our analysis to the 1000 genes with the largest variance and 193 miRNAs with no missing values for 500 samples available for both data types. The data matrices are scaled using the Frobenius norm, \( X_m/\|X_m\|_F \). In addition, information on disease-specific survival, estrogen receptor and HER2 status is available. The outline of our analysis is as follows: First, the numbers of joint and individual clusters are chosen, and then the clusters are explored with respect to survival time and estrogen receptor status.

We determine the numbers of joint, gene expression-specific and miRNA-specific clusters, \( K, K_1 \) and \( K_2 \) by the procedure described in Section 2.4. Figure 1 displays density plots and qq-plots for the first four component scores of the concatenated matrix \( X \), while the density and qq-plots from the fifth to the eighth component can be found in the Supplementary material. As the component number increases, the component scores become closer to being normally distributed. Based on the qq-plots, the first, second and third joint components are considered not normally distributed, while the distribution of the fourth component scores and further do
not deviate significantly from normality, such that $E = 3$. The density and qq-plots for the first to the eighth component scores of the gene expression and miRNA alone are found in the Supplementary material.

The choice of three joint components sets the upper limit for the number of relevant components in individual data sets. The qq-plots of the first three component scores of the original expression data suggest that the first, second and third component deviate significantly from normality, while for the fourth component and upwards there seem to be no significant deviations. For miRNA, only the first two components deviate from normality, while the third to the eight component do not. We therefore determine the number of relevant components in the gene expression and the miRNA data to be $E_1 = 3$ and $E_2 = 2$, respectively.

With the numbers of relevant components $E = 3, E_1 = 3$ and $E_2 = 2$, the number of clusters are given by Equation (2.3) as

$$K = 3, \quad K_1 = 2, \quad K_2 = 1,$$

meaning three joint clusters, two expression-specific clusters and no miRNA-specific clusters.

Figure 2 displays the first and second joint component scores colored according to the three joint clusters. The first component separates out the green cluster, while the second component discriminates between the blue and red cluster. A negative or positive Estrogen Receptor (ER) status for the patient is marked by filled and open circles, respectively, revealing that most ER-negative samples coincide with the green cluster.

Figure 3 displays the first and second principal component scores of the original gene expression and miRNA data. In the upper panel of Figure 3, the observations are colored according to the three joint clusters, showing that the joint clusters are reflected in both the original data sources. While in the lower panel of Figure 3, the observations are colored according to the two clusters specific to expression, demonstrating that the expression-specific clusters are independent of miRNA.

We further compare the survival of the different joint and gene expression-specific clusters independently through Kaplan-Meier estimates. However, due to the low number of events neither the three joint groups nor the two gene expression-specific groups are significantly different. But these observations suggest that there may be two independent mechanisms present, one jointly for gene expression and miRNA, and one specific to gene expression.

5. Discussion

With Joint and Individual Clustering it is possible to decompose data into common and data type-specific components giving rise to meaningful clusters. Our example with gene expression and miRNA from breast cancer tissue shows that in addition to cancer subtypes found jointly in different data types, there may exist independent groups only found in a single data type.

The selection of the number of clusters is a crucial step in the analysis and has proved difficult due to the large number of cluster configurations and the combination of data types with different characteristics. To be able to analyze clusters that are not well-separated, our approach assumes the noise to be normally distributed. This assumption is reasonable for data such as gene expression, miRNA and proteins, but not for methylation data or copy number variants. Future work needs to also incorporate such types of data, for instance by changing the distributional assumption for the noise.
REFERENCES

6. SUPPLEMENTARY MATERIAL

Supplementary material, containing R code and additional plots, is available online at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENT

The authors would like to thank J.S. Marron for guidance regarding the JIVE algorithm. Funding for the project was provided by the Norwegian Cancer Society under award 744088. Conflict of Interest: None declared.

REFERENCES

Arabie, P. and Hubert, L. (1996). Advances in cluster analysis relevant to marketing research. In: From Data to Knowledge. Springer, pp. 3–19.

Cancer Genome Atlas Network. (2012). Comprehensive molecular portraits of human breast tumours. Nature 490(7418), 61–70.

Chalise, P., Koestler, D. C., Bimali, M., Yu, Q. and Fridley, B. L. (2014). Integrative clustering methods for high-dimensional molecular data. Translational cancer research 3(3), 202.

Chang, W. (1983). On using principal components before separating a mixture of two multivariate normal distributions. Applied Statistics, 267–275.

Deun, K. Van, Smilde, A. K., van der Werf, M. J., Kiers, A. L. and Mecheelen, I. Van. (2009). A structured overview of simultaneous component based data integration. BMC bioinformatics 10(1), 246.

Deun, K. Van, van den Berg, T. Wilderjansand R., Antoniadis, A. and Mecheelen, I. Van. (2011). A flexible framework for sparse simultaneous component based data integration. BMC bioinformatics 12(1), 448.

Ding, C. and He, X. (2004). K-means clustering via principal component analysis. In: Proceedings of the twenty-first international conference on Machine learning. pp. 29–41.

Hamerly, G. and Elkan, C. (2003). Learning the k in k-means. In: Advances in Neural Information Processing Systems, Volume 3. pp. 281–288.

Hellton, K. H. and Thoresen, M. (2014). Asymptotic distribution of principal component scores for pervasive, high-dimensional eigenvectors. arXiv preprint arXiv:1401.2781.

Kaufman, P. J. L. and Rousseeuw. (2009). Finding groups in data: an introduction to cluster analysis. John Wiley & Sons.

Kirk, P., Griffin, J. E., Savage, R. S., Ghahramani, Z. and Wild, D. L. (2012). Bayesian correlated clustering to integrate multiple datasets. Bioinformatics 28(24), 3290–3297.

Kormaksson, M., Booth, J. G., Figueroa, M. E. and Melnick, A. (2012). Integrative model-based clustering of microarray methylation and expression data. The Annals of Applied Statistics 6(3), 1327–1347.
REFERENCES

LOCK, E. F. AND DUNSON, D. B. (2013). Bayesian consensus clustering. Bioinformatics 29(20), 2610–2616.

LOCK, E. F., Hoadley, K. A., Marron, J. S. AND NOBEL, A. B. (2013). Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. The annals of applied statistics 7(1), 523–542.

LÖFSTEDT, T., HOFFMAN, D. AND TRYGG, J. (2013). Global, local and unique decompositions in OnPLS for multiblock data analysis. Analytica chimica acta 791, 13–24.

RAY, P., ZHENG, L., LUCAS, J. AND CARIN, L. (2014). Bayesian joint analysis of heterogeneous genomics data. Bioinformatics 30(10), 1370–1376.

SCHOUTEDEN, M., VAN DEUN, K., PATTYN, S. AND VAN MECHELEN, I. (2013). SCA with rotation to distinguish common and distinctive information in linked data. Behavior research methods 45(3), 822–833.

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(22), 2906–2912.

SHEN, R., WANG, S. AND MO, Q. (2013). Sparse integrative clustering of multiple omics data sets. The annals of Applied statistics 7(1), 269–294.

TIBSHIRANI, R. AND WALTHER, G. (2005). Cluster validation by prediction strength. Journal of Computational and Graphical Statistics 14(3), 511–528.

TIPPING, M. E. AND BISHOP, C. M. (1999). Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 61(3), 611–622.

YEUNG, K. Y. AND RUZZO, W. L. (2001). Principal component analysis for clustering gene expression data. Bioinformatics 17(9), 763–774.

ZHA, H., HE, X., DING, C., GU, M. AND SIMON, H. D. (2001). Spectral relaxation for k-means clustering. In: Advances in Neural Information Processing Systems, Volume 1. pp. 1057–1064.

ZHANG, S., LIU, C., LI, W., SHEN, H., LAIRD, P. W. AND ZHOU, X. J. (2012). Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. Nucleic acids research 40(19), 9379–9391.

[Received September 5, 2014; revised July 25, 2015; accepted for publication ———]
Fig. 1. Density plots and normal quantile-quantile plots for the four first joint component scores.

Fig. 2. The first and second joint component scores colored according to the three joint clusters. The Estrogen Receptor (ER) status for the patient, either negative or positive, is marked as open or filled circle, respectively.

Fig. 3. The first and second principal component of the original expression and miRNA data, colored according to the three joint clusters in the upper panel and the two clusters specific to expression in the lower panel.

Table 1. Mean precision of the estimated cluster assignments for 100 simulations when the numbers of clusters are known. The percentage of times the numbers of clusters are correctly estimated (pct. of corr.) for 100 simulations. For all simulations, \( c = 80 \).

| Setting I | iCluster | JIC - joint | JIC - X_1 | JIC - X_2 | JIC - X_3 |
|-----------|----------|-------------|-----------|-----------|-----------|
| True K    | 5        | 1           | 1         | 1         |           |
| Precision (SD) | 0.998 (0.004) | 0.985 (0.01) | –         | –         | –         |
| Pct. of corr. K | 97%       | 96%         | 95%       | 98%       |           |
| Setting II| iCluster | JIC - joint | JIC - X_1 | JIC - X_2 | JIC - X_3 |
| \( c_1=c_2=c_3=32 \) | True K    | 5           | 2         | 2         | 2         |
| Precision (SD) | 0.415 (0.077) | 0.933 (0.097) | 0.950 (0.112) | 0.791 (0.165) | 0.874 (0.168) |
| Pct. of corr. K | 89%       | 90%         | 88%       | 88%       |           |
| \( c_1=c_2=32, c_3=16 \) | Precision (SD) | 0.619 (0.026) | 0.944 (0.089) | 0.993 (0.036) | 0.805 (0.135) | 1 (-) |
| Pct. of corr. K | 94%       | 98%         | 88%       | 86%       |           |
| \( c_1=32, c_2=c_3=16 \) | Precision (SD) | 0.807 (0.049) | 0.909 (0.101) | 0.711 (0.070) | 1 (-) | 1 (-) |
| Pct. of corr. K | 88%       | 87%         | 91%       | 89%       |           |
| \( c_1=c_2=c_3=16 \) | Precision (SD) | 1 (-) | 1 (-) | 1 (-) | 1 (-) | 0.999 (0.003) |
| Pct. of corr. K | 53%       | 49%         | 50%       | 49%       |           |