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

Kristoffer Hellton, Magne Thoresen
Oslo Center for Biostatistics and Epidemiology, Department of Biostatistics, University of Oslo, Norway
Department of Mathematics, University of Oslo, Norway
kristohh@math.uio.no

Abstract

It is now possible to measure a range of genomic data types or data layers in a single tissue sample, e.g. gene expression, copy number variations, miRNA or mutation markers, and clustering of patient samples benefits from an integrative approach using all available data types. Earlier methods, however, are restricted to only allowing a joint cluster structures, equal in all data layers. We present a clustering extension of the Joint and Individual Variance Explained (JIVE) algorithm enabling construction of both joint and data type-specific clusters simultaneously (Hellton and Thoresen, 2016).

Method

Let $X_1, \dots, X_M$ be M genomic data types measured on the same $j = 1, \dots, n$ patients, concatenated into a single matrix

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_M \end{bmatrix},$$

of dimension $(p_1 + \cdots + p_M) \times n$. Joint and Individual Clustering (JIC) assumes latent common and data type-specific cluster indicators, Z and $Z_1, \dots, Z_M$,

$$X_1 = W_1 Z + V_1 Z_1 + \varepsilon_1,$$
$$\vdots$$
$$X_M = W_M Z + V_M Z_M + \varepsilon_M,$$

with $\epsilon_m \sim N(0, \sigma_m^2 I)$ for $m = 1, \dots, M$ and joint and individual loading matrices $W = [W_1, \dots, W_M]^T$ and $V_1, \dots, V_M$.

For given numbers of clusters K and $K_1, \dots, K_M$, the clusters are estimated by a two-stage procedure based on the singular value decomposition (SVD):

1. Estimate the joint and individual matrices, WZ and $V_m Z_m$ using the JIVE algorithm (Lock et al. 2013). Initialize $X^{\text{JOINT}} = X$ and repeat the following steps until convergence,
   - set WZ to be the SVD of X^{JOINT} with rank $r = K - 1$,
   - for $m = 1, \dots M$, set $V_m Z_m$ to be the SVD of $X_m^{\text{INDIVID}}(I - ZZ^T)$ with rank $r_m = K_m - 1$, where $X_m^{\text{INDIVID}} = X_m - W_m Z$,
   - concatenate the matrices $X^{\text{JOINT}} = [X_1 - V_1 Z_1, \dots, X_M - V_M^T Z_M]^T$.

   After convergence, Z is given by the r first right singular vectors of X^{JOINT} and Z_m is given by the r_m first right singular vectors of X_m^{INDIVID} for $m = 1, \dots, M$.
2. Cluster the rows of Z^T into K groups and the rows of Z_m^T into K_m groups for $m = 1, \dots, M$ using k-means clustering.

Selecting the numbers of clusters

The numbers of clusters, K and $K_1, \dots, K_M$, are chosen by assessing the number of non-normally distributed components in the concatenated and original data types, separately, under the assumption of normally distributed errors:

1. Assess the number of relevant components E in X: check the normality of the ith 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 $m = 1, \dots, M$, check the normality of the ith 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 $E, E_1, \dots, E_M$, calculate the ranks $r, r_1, \dots, r_m$ by

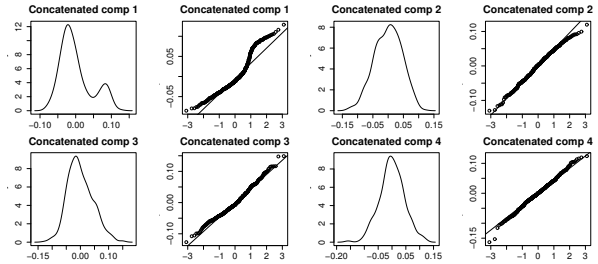
$$r = \frac{E_1 + \cdots + E_M - E}{M - 1}, \qquad r_m = E_m - r, \quad m = 1, \dots, M.$$

   such that the numbers of clusters are $K = r + 1$ and $K_m = r_m + 1$.

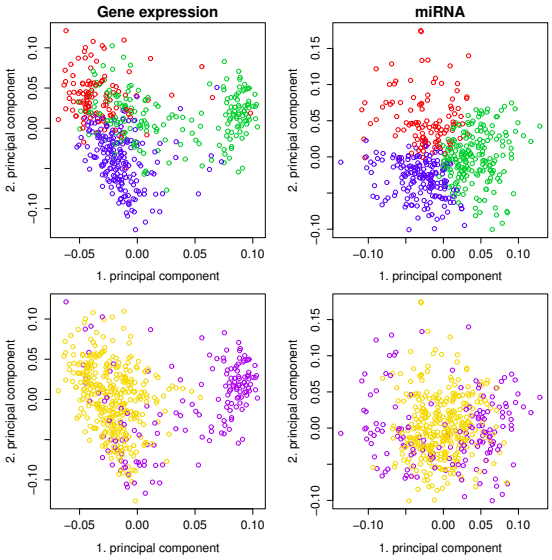
With the connection between the number of clusters and the rank of the latent structures (Ding and He, 2004), we check whether a new cluster is separated out in each added component until the first component with only a single cluster is found. When assuming normally distributed noise, it is possible to identify overlapping clusters by evaluating density and normal quantile-quantile plots. If the distribution of component scores deviates from normality, the component does not represent noise and clusters should be present.

Example: gene expression and miRNA

The method is illustrated using data from The Cancer Genome Atlas (TCGA); gene expression ($p_1 = 1000$) and miRNA ($p_2 = 193$) measured in 500 breast cancer tissue samples. The density and qq-plots for the concatenated data (shown below) suggest that the three first components are not normally distributed, while the original expression and miRNA data suggest that the three and two first components deviate from normality, respectively.



This results in three joint clusters, $K = 3$, two expression-specific clusters, $K_1 = 2$, and no miRNA-specific clusters, $K_2 = 1$.



Coloring the observations according to the three joint clusters (upper panels) shows that the joint clusters are reflected in both the original data sources, while coloring the observations according to the two expression-specific clusters (lower panels) demonstrates how they only reflect the structure of expression data and not the structure of miRNA. Differences in survival between the different cluster configurations were not significant due to a low number of events.

References

K. H. Hellton and M. Thoresen (2016). *Integrative clustering of high-dimensional data with joint and individual clusters*, Biostatistics, kxw005.

E. F. Lock and K. A. Hoadley and J. S. Marron and A. B. Nobel (2013). *Joint and individual variation explained (JIVE) for integrated analysis of multiple data types*, The Annals of Applied Statistics, Volume 7, pages 523–542.

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

Conclusion

The Joint and Individual Clustering makes it possible to decompose data into common and data type-specific clusters following the JIVE algorithm. Example data from the Cancer Genome Atlas consisting of gene expression and miRNA from breast cancer tissues illustrates how there may exist independent groups only found in a single data type, in addition to cancer subtypes found jointly in different data types.