iCluster Bayes

Stephen Coleman

24/10/2019

Notation

N	The number of samples in the data.
L	The number of datasets present.
p_l	The number of measurements for each sample in the lth dataset.
K_l	The number of components present in the lth dataset.
$X = (X_1, \ldots, X_l)$	The datasets.
$X_l = (X_{l1}, \ldots, X_{lN})$	The observed data for the lth dataset.
$c = (c_1, \ldots, c_L)$	The membership vectors for each dataset (our latent variable).
$c_l = (c_{l1}, \ldots, c_{lN})$	The context-specific component membership.
$C = (C_1, \ldots, C_N)$	The global allocation vector.
$\pi_l = (\pi_{l1,} \ldots, \pi_{lK_l})$	The mixture weights in the lth context.

If $K_1 = \ldots = K_L$ then we use K as the number of components in each context. We treat p_l in the same way. I denote abbreviations or terms that will be used in place of another in the format "[Full name] ([name hereafter])".

Intro

The Bayesian latent variable model (iCluster Bayes) proposed by Mo et al. (2017) is an extension of the Gaussian latent variable model proposed by Shen, Olshen, and Ladanyi (2009). This model has already been extended to incorporate feature selection (Shen, Wang, and Mo 2013). This model maps from the data, X, to a low dimensional subspace, Z. This is map from the L high dimensional spaces to a single $N \times K$ space, where the ith sample has an associated vector of values $z_i = (z_{i1}, \ldots, z_{iK}) \forall i \in [1, N]$. z_i is a continuous variable and $z_i \sim \mathbf{MVN}(\mathbf{0}, \mathbf{I}_K)$. In other models the data is mapped to a $N \times K$ space of probabilities before being assigned to specific clusters. This space, Z, does not consist of probabilities. The model then uses k-means clustering on this space where $k = K + 1$.

Model

The model is based upon factor analysis. COnsider the single dataset case initially and let X be in the form of $[measurements \times samples]$, thus it is a $p \times N$ matrix. Then we model this using a factor analysis:

$$X = LF + \epsilon.$$

Here the loadings matrix, L, is a $p \times (K + 1)$ matrix and the factors, F, are encoded in a $(K + 1) \times N$ matrix.

$$L = \begin{bmatrix} l_{10} & l_{11} & \cdots & l_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ l_{p0} & l_{p1} & \cdots & l_{pK} \end{bmatrix}; F = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ z_{11} & z_{12} & \cdots & z_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ z_{K1} & z_{K2} & \cdots & z_{KN} \end{bmatrix}.$$

iCluster aims to reduce the dimensionality of the problem. They change the framing of the problem. Consider a single row of X, associated with the jth feature (for example a gene if X consists of gene expression data) and denote this X_j. Similarly let L_j be the row of the loadings matrix associated with the jth feature. Now let us say:

$$X_j^T = F^T L_j^T + \epsilon$$

The next step is the inclusion of a specific $(K+1) \times (K+1)$ sparsity inducing matrix, Γ_j, which has the form $diag(1, \gamma_j, \ldots, \gamma_j)$; this will allows to discard a subset of features as uniformative (George and McCulloch 1997). The inclusion of a constant 1 in both Γ and F allows an intercept value of l_{j0}. γ_j is a binary variable. If the value of a given loading, l_j, is small enough that ignoring it is preferable, then the γ_j takes a value of 0; otherwise it is 1. In this case our model takes the form:

$$X_j^T = F^T \Gamma_j L_j^T + \epsilon.$$

Consider an example where $N = 4$ and $K = 3$. Then:

$$\begin{aligned}
\begin{bmatrix} x_{j1} \\ x_{j2} \\ x_{j3} \\ x_{j4} \end{bmatrix} &= \begin{bmatrix} 1 & z_{11} & z_{12} \\ 1 & z_{21} & z_{22} \\ 1 & z_{31} & z_{32} \\ 1 & z_{41} & z_{42} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \gamma_j & 0 \\ 0 & 0 & \gamma_j \end{bmatrix} \begin{bmatrix} l_{j0} \\ l_{j1} \\ l_{j2} \end{bmatrix} \\
&= \begin{bmatrix} 1 & z_{11}\gamma_j & z_{12}\gamma_j \\ 1 & z_{21}\gamma_j & z_{22}\gamma_j \\ 1 & z_{31}\gamma_j & z_{32}\gamma_j \\ 1 & z_{41}\gamma_j & z_{42}\gamma_j \end{bmatrix} \begin{bmatrix} l_{j0} \\ l_{j1} \\ l_{j2} \end{bmatrix} \\
&= \begin{bmatrix} l_{j0} + \gamma_j(z_{11}l_{j1} + z_{12}l_{j2}) \\ l_{j0} + \gamma_j(z_{21}l_{j1} + z_{22}l_{j2}) \\ l_{j0} + \gamma_j(z_{31}l_{j1} + z_{32}l_{j2}) \\ l_{j0} + \gamma_j(z_{41}l_{j1} + z_{42}l_{j2}) \end{bmatrix}
\end{aligned}$$

One can see that if $\gamma_j = 0$ that the feature is loaded only through the intercept value of l_{j0}, whereas if $\gamma_j = 1$ the remaining terms are included. I believe the entire dataset is then modelled as:

$$X^T = \begin{bmatrix} l_{10} + \gamma_1 \sum_{k=1}^K l_{1k} z_{1k} & \cdots & l_{p0} + \gamma_p \sum_{k=1}^K l_{pk} z_{1k} \\ \vdots & \ddots & \vdots \\ l_{10} + \gamma_1 \sum_{k=1}^K l_{1k} z_{Nk} & \cdots & l_{p0} + \gamma_p \sum_{k=1}^K l_{pk} z_{Nk} \end{bmatrix}$$

Continuous data

OK there's some issues around this point in the original paper []. They use a slightly different notation than I do, letting $\beta_j = L_j \forall j \in [1, p]$. They let the priors be based around a vector β_0 which they never define. I shall assume it's some prior parameter (possibly the mean of the entire dataset), but it is possible that I have misunderstood it.

In the case that the data is continuous, then the following priors are assumed:

$$L_j \sim \mathbf{MVN}\left(L_0, \Sigma_0\right), \sigma_j^2 \sim \text{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right), \gamma_j \sim \text{Bernoulli}\left(q\right)$$

From these a posterior distribution can be derived:

References

George, Edward I, and Robert E McCulloch. 1997. "Approaches for Bayesian Variable Selection." *Statistica Sinica*, 339–73.

Mo, Qianxing, Ronglai Shen, Cui Guo, Marina Vannucci, Keith S Chan, and Susan G Hilsenbeck. 2017. "A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data." *Biostatistics* 19 (1): 71–86. https://doi.org/10.1093/biostatistics/kxx017.

Shen, Ronglai, Adam B Olshen, and Marc Ladanyi. 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): 2906–12.

Shen, Ronglai, Sijian Wang, and Qianxing Mo. 2013. "Sparse Integrative Clustering of Multiple Omics Data Sets." *The Annals of Applied Statistics* 7 (1): 269.