A joint mixture model for clustering genes from Gaussian and beta distributed data

TAMPERE UNIVERSITY OF TECHNOLOGY

Xiaofeng Dai, Harri Lähdesmäki & Olli Yli-Harja

Department of Signal Processing, Tampere University of Technology, Tampere, Finland

Motivation

- Overall motivation: building a single and unified probabilistic modeling framework to cluster genes from different data sources, which can provide a more efficient use of multiple data sources than methods that analyze different data separately.
- Motivation in this paper: building a mixture model based clustering method which can cluster genes jointly from protein-DNA binding probabilities and gene expression data.

Methods

- The mixture model is defined as

$$f(\mathbf{x_j}|\theta) = \sum_{i=1}^{g} \pi_i f_i^{(g)}(\mathbf{x_j}|\theta_i) \qquad (1)$$

- BGMM is built with the assumption that, for each component i, the beta distributed and Gaussian distributed data are independent.
- PDF of beta and Gaussian distribution are

$$f_i(\mathbf{y}|\theta_{1i}) = \prod_{u=1}^{p_1} \frac{y_u^{\alpha_{iu}-1}(1-y_u)^{\beta_{iu}-1}}{B(\alpha_{iu},\beta_{iu})}, \qquad (2)$$

$$f_i(\mathbf{z}|\theta_{2i}) = \frac{1}{(2\pi)^{\frac{p_2}{2}}|V|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{z}-\mu_i)^T V^{-1}(\mathbf{z}-\mu_i)\right), \qquad (3)$$

- Three EM algorithms are derived for this problem, EM_s, EM_a and EM_h.
  - EM_a: both BMM and GMM maximize the complete data log-likelihood

$$\log L_c(\theta) = \sum_{j=1}^{n}\sum_{i=1}^{g} \chi(c_j = i)\log(\pi_i f_i(\mathbf{x}_j|\theta_i)), \qquad (4)$$

  - EM_s: both BMM and GMM maximize Q, the expectation of Eq. 4

$$Q(\Theta;\Theta^{(m)}) = E_{\Theta^{(m)}}(\log L_c|X) = \sum_{j=1}^{n}\sum_{i=1}^{g} \tau_{ji}^{(m)} \log(\pi_i f_i(x_j;\theta_i)) \qquad (5)$$

  - EM_h: BMM maximizes Eq. 4, and GMM maximize Eq. 5.
- Parameter updates
  - EM_a: α_{iu}'s and β_{iu}'s are estimated with 'betafit'; μ_{iv}'s and σ_v's are calculated by

$$\hat{\mu}_i^{(m+1)} = \sum_{j\in I_i^{(m)}} z_{jv}^{(m)}/n_i^{(m)}, \qquad (6)$$

$$\hat{\sigma}_v^{2,(m+1)} = \sum_{j\in I_i^{(m)}}\sum_{i=1}^{g}(z_{jv}-\mu_{iv}^{(m)})^2/n. \qquad (7)$$

  - EM_s: α_{iu}'s and β_{iu}'s are estimated using Newton-Raphson method

$$\theta_{1i}^{(m+1)} = \theta_{1i}^{(m)} - H^{-1}(\theta_{1i}^{(m)})\nabla_{\theta_{1i}}\mathcal{L}(\theta_{1i}^{(m)}) \qquad (8)$$

μ_{iv}'s and σ_v^2's are estimated by iterating over

$$\hat{\mu}_i^{(m+1)} = \sum_{j=1}^{n} \tau_{ji}^{(m)} z_{jv}/\sum_{j=1}^{n} \tau_{ji}^{(m)}, \qquad (9)$$

$$\hat{\sigma}_v^{2,(m+1)} = \sum_{j=1}^{n}\sum_{i=1}^{g} \tau_{ji}^{(m)}(z_{jv}-\mu_{iv}^{(m)})^2/n. \qquad (10)$$

  - EM_h: α_{iu}'s and β_{iu}'s are updated with 'betafit'; μ_{iv}'s and σ_v^2's are estimated according to Eq. 9 and Eq. 10.
  - π: updated by $\pi_i^{(m+1)} = \sum_{j=1}^{n} \tau_{ji}^{(m)}/n$.
- Model selection: BIC, ICL-BIC, AIC, and AIC3 are compared in each model, based on which each best criterion is chosen.

Results

- Scoring system: named 'E' score, used for evaluating the accuracy of the tested models

$$e_j(r) = \begin{cases} 1 \; if \;\; \hat{z}_{ji} = 1 \;\; and \;\; r_i = T_j \\ 0 \; otherwise \end{cases}$$

$$E = \max_{r\in R}\sum_{j=1}^{n} e_j(r)/n$$

$$R = \{r = (r_1,\ldots,r_{\hat{g}}) : \forall i \neq j \; r_i \neq r_j; \; r_i \in \{1,\ldots,\max\{\hat{g},g\}\}\}.$$

- Data sets: good beta (gB), good Gaussian (gG), bad beta (bB), bad Gaussian with respect to close means (bG_m), bad Gaussian with respect to large variances (bG_v)
- Comparison among three EM's: The accuracy of BGMM is compared with BMM and GMM, whose results are illustrated below:

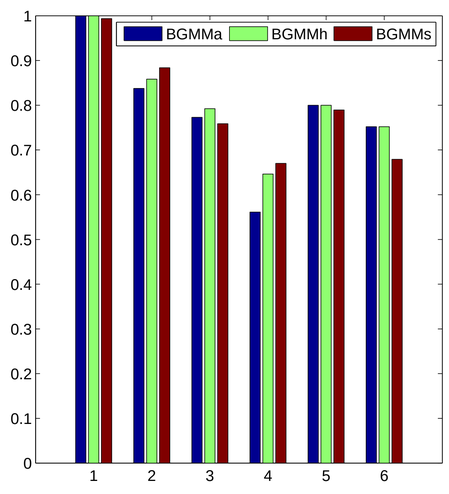

FIGURE 1: Comparison of different EM's. 1: gB+gG, 2: bB+gG, 3: gB+bG_m, 4: bB+bG_m, 5: gB+bG_v, 6: bB+bG_v

- Comparison of BGMM with BMM and GMM: The accuracy of BGMM is compared with BMM and GMM, whose results are illustrated below:

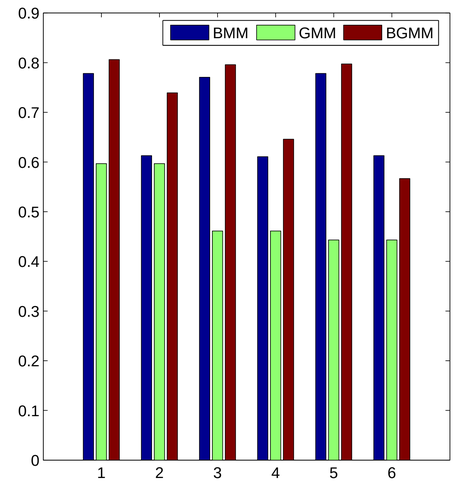

FIGURE 2: Performance test of BGMM. 1: gB+gG, 2: bB+gG, 3: gB+bG_m, 4: bB+bG_m, 5: gB+bG_v, 6: bB+bG_v

- Criteria selection: we summed up the number of hits of the correct number of clusters for each data combination in both simulations, according to which ICL works well in $BGMM_s$, AIC is proposed for $BGMM_a$ and $BGMM_h$.
- Real case study: applied to 673 mouse genes
  - GO validation: clusterings by BGMM have lower p-values compared with BMM and GMM.
  - Cluster analysis

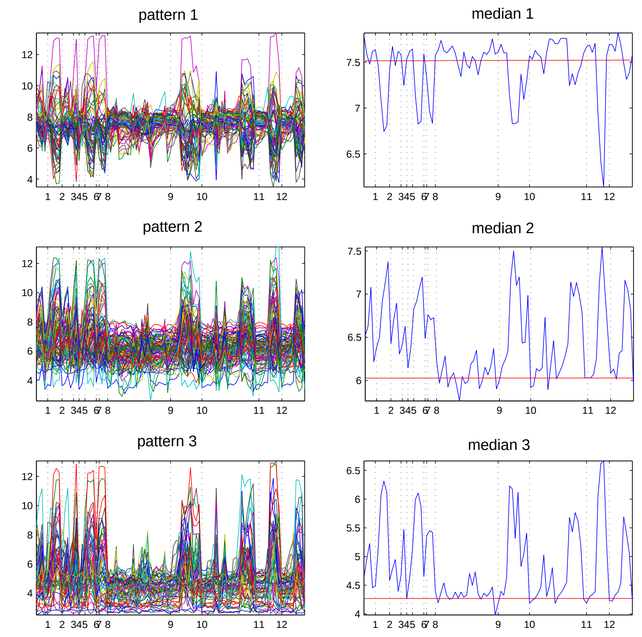

FIGURE 3: Expression patterns of clusters obtained from BGMM.

Conclusions & Future Directions

- BGMM can be applied to any problems whose data can be modeled as beta and Gaussian distribution, respectively.
- BGMM can be easily extended to combine data of any other parametric distributions in principle.
- Future work
  - Integrate more data sources into this framework to improve clustering accuracy.
  - develop similar joint model but with different parametric distributions that are needed in other problems.
  - Use additional prior to stratify the joint mixture model, which can utilize any type of information.