13.2.2 Biomarker Selection in Mixture Models

We now address variable selection in a different modeling context, that is, mixture models for pattern recognition. We treat in particular the unsupervised framework, known in the statistical literature as clustering, and then describe an adaptation to the simpler supervised framework, known as discriminant analysis. For both model formulations, we borrow ideas from the linear settings treated in the previous section. For example, a latent binary vector γ is introduced for variable selection, and stochastic search MCMC techniques are used to explore the space of variable subsets. However, building a variable selection mechanism into mixture models is more challenging than the linear settings. In clustering, for example, there is no observed response to guide the selection, and the elements of the matrix $\mathbf{X}$ are viewed as random variables. The inclusion of the latent indicators into the models, therefore, cannot be done like in the linear modeling context, where γ is used to induce mixture priors on regression coefficients.

An approach to variable selection for model-based clustering was put forward by Tadesse et al. (2005), who formulated the clustering in terms of a finite mixture of Gaussian distributions with an unknown number of components and then introduced latent variables to identify discriminating variables. The authors used a reversible jump MCMC technique to allow for the creation and deletion of clusters. A similar model was considered by Raftery and Dean (2006). Kim et al. (2006) proposed an alternative modeling approach that uses infinite mixture models via Dirichlet process priors. Hoff (2006) adopted a mixture of Gaussian distributions where different clusters are identified by mean shifts and Bayes factors are computed to identify discriminating variables. This method allows separate subsets of variables to discriminate different groups of observations.

In the finite mixture model formulation of Tadesse et al. (2005), the data are viewed as coming from a mixture of distributions

$$f(\mathbf{x}_i|\mathbf{w}, \phi) = \sum_{k=1}^{K} w_k f(\mathbf{x}_i|\phi_k), \tag{13.9}$$

where $f(\mathbf{x}_i|\phi_k)$ is the density of sample $\mathbf{x}_i$ from group k and $\mathbf{w} = (w_1, \ldots, w_K)^T$ are the cluster weights ($\sum_k w_k = 1, w_k \geq 0$); see McLachlan and Basford (1988). Here K is assumed finite but unknown. Latent variables $\mathbf{c} = (c_1, \ldots, c_n)^T$, with $c_i = k$ if the i-th sample comes from group k, are introduced to identify the cluster from which each observation is drawn. The sample allocations, c_i, are assumed to be independently and identically distributed with probability mass function $p(c_i = k) = w_k$. We assume that the mixture

distributions are multivariate normal with component parameters $\phi_k = (\mu_k, \Sigma_k)$. Thus, for sample i, we have

$$\mathbf{x}_i | c_i = k, \mathbf{w}, \phi \sim \mathcal{N}(\mu_k, \Sigma_k). \tag{13.10}$$

For variable selection, a latent binary vector γ is used to identify the discriminating variables. More specifically, variables indexed by a $\gamma_j = 1$, denoted $\mathbf{X}_{(\gamma)}$, define the mixture distribution, whereas variables indexed by $\gamma_j = 0$, $\mathbf{X}_{(\gamma^c)}$, favor one multivariate normal distribution across all samples. The distribution of sample i is then given by

$$\begin{aligned} \mathbf{x}_{i(\gamma)} | c_i = k, \mathbf{w}, \phi, \gamma &\sim \mathcal{N}(\mu_{k(\gamma)}, \Sigma_{k(\gamma)}) \\ \mathbf{x}_{i(\gamma^c)} | \psi, \gamma &\sim \mathcal{N}(\eta_{(\gamma^c)}, \Omega_{(\gamma^c)}), \end{aligned} \tag{13.11}$$

where $\psi = (\eta, \Omega)$.

Priors on γ can be specified similarly to what was discussed for the linear settings, via MRF distributions. For the vector of component weights, a symmetric Dirichlet prior can be specified. For the unknown number of components, K, a truncated Poisson or a discrete Uniform prior on $[1, \ldots, K_{\max}]$, where $K_{\max}$ is chosen arbitrarily large, are suitable choices. An efficient sampler can be implemented by working with a marginalized likelihood where the model parameters are integrated out. The integration is facilitated by taking conjugate Normal-Wishart priors on both ϕ and ψ. Some care is needed in the choice of the hyper-parameters. In particular, the variance parameters need to be specified within the range of variability of the data. The MCMC procedure is described in Tadesse et al. (2005) and requires a sampler that jumps between different dimensional spaces, generalizing the reversible jump approach of Richardson and Green (1997).

Discriminant Analysis

We now show an adaptation of the method to the simpler supervised setting, where, in addition to the observed vectors $\mathbf{x}_i$'s, the number of groups G and the classification labels c_i's are also available and where the aim is to derive a classification rule that will assign further cases to their correct groups. When the distribution of $\mathbf{X}$ conditional on the group membership is assumed normal, then this statistical methodology is known as discriminant analysis.

In discriminant analysis, given the selected variables, the predictive distribution of a new observation $\mathbf{x}^f$ is used to classify every new sample into one of the possible G groups. This distribution is a multivariate T-student, see Brown (1993) among others. The probability that a future observation, given

the observed data, belongs to the group k is then given by

$$\pi_k(c^f|\mathbf{X}) = p(c^f = k|\mathbf{x}^f, \mathbf{X}), \tag{13.12}$$

where c^f is the group indicator of $\mathbf{x}^f$. By estimating the prior probability that one observation comes from group k as $\hat{\pi}_k = n_k/n$, the previous distribution can be written in closed form as

$$\pi_k(c^f|\mathbf{X}) = \frac{p_k(\mathbf{x}^f)\hat{\pi}_k}{\sum_{i=1}^{G} p_i(\mathbf{x}^f)\hat{\pi}_i},$$

where $p_k(\mathbf{x}^f)$ indicates the predictive T-student distribution. A new observations is then assigned to the group with the highest posterior probability.

As in the clustering setting, we introduce a latent binary vector γ to perform the selection. As done by Raftery and Dean (2006), extending the approach of Tadesse et al. (2005) to avoid any independence assumptions, the following likelihood can be used to separate the discriminant variables from the noisy ones as

$$L(X, c; \cdot) = \prod_{i=1}^{n} p(\mathbf{x}_{i(\gamma^c)}|\mathbf{x}_{i(\gamma)}) \prod_{k=1}^{G} w_k^{n_k} \prod_{j=1}^{n_k} p_k(\mathbf{x}_{j(\gamma)}).$$

The first factor of the likelihood refers to the non-important variables, whereas the second is formed by variables able to classify observations into the correct groups. Under the normality assumption, the likelihood becomes

$$\prod_{i=1}^{n} N_{|\gamma^c|}(\mathbf{x}_{i(\gamma^c)} - \beta\mathbf{x}_{i(\gamma)}; \eta_{(\gamma^c)}, \Sigma_{(\gamma^c)}) \prod_{k=1}^{G} w_k^{n_k} \prod_{j=1}^{n_k} N_{|\gamma|}(\mathbf{x}_{j(\gamma)}; \mu_{k(\gamma)}, \Sigma_{k(\gamma)}).$$

As in the unsupervised case, conjugate Normal-Wishart priors can be specified on the model parameters. Again, the corresponding MCMC algorithm benefits of this parametrization, because it is possible to integrate out means and variances and design Metropolis steps that depend only on the selected and proposed variables.

An Application to Microarray Data

We show an application from Stingo and Vannucci (2011) on the widely used leukemia data of Golub et al. (1999) that comprise a training set of 38 patients and a validation set of 34 patients. An MRF prior on γ is used to capture knowledge on the gene network structure, as extracted from KEGG. Note that some of the genes do not have neighbors. The training set consists of bone marrow samples obtained from acute leukemia patients, whereas the validation set consists of 24 bone marrow samples and 10 peripheral blood samples.