Supplemental material S1: Integrative mixture of experts to combine clinical factors and gene markers

1 Methods

1.1 Parameter estimation via the EM algorithm

To apply the EM algorithm to the ME architecture, we introduce the indicator variables ζ_{hj}, where ζ_{hj} is one if y_j belongs to the h^{th} expert or zero otherwise.

The complete data likelihood for $\boldsymbol{\Psi}$ is given by

$$\log L_c(\boldsymbol{\Psi}) = \sum_{j=1}^{n}\sum_{h=1}^{H} \zeta_{hj}\{\log \boldsymbol{\pi}_h + \log f_h^G(\boldsymbol{w}_j; \boldsymbol{\alpha}_h) + \log f_h^E(y_j|\boldsymbol{w}_j; \boldsymbol{\beta}_h)\} \tag{1}$$

When we apply the EM algorithm to train the ME architecture, the E-step calculates the Q-function on the $(k+1)^{th}$ iteration as

$$\begin{aligned} Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)}) &= \mathbb{E}_{\boldsymbol{\Psi}^{(k)}}\{\log L_c(\boldsymbol{\Psi})|\boldsymbol{y}, \boldsymbol{w}\} \\ &= \sum_{j=1}^{n}\sum_{h=1}^{H} \mathbb{E}_{\boldsymbol{\Psi}^{(k)}}(\zeta_{hj}|\boldsymbol{y}, \boldsymbol{w})\{\log \boldsymbol{\pi}_h + \log f_h^G(\boldsymbol{w}_j; \boldsymbol{\alpha}_h) \\ &+ \log f_h^E(y_j|\boldsymbol{w}_j; \boldsymbol{\beta}_h)\}. \end{aligned} \tag{2}$$

As (2) is linear in ζ_{hj}, the E-step replaces ζ_{hj} in (1) by its current conditional expectation $\tau_{hj}^{(k)}$ given $y_j, \boldsymbol{w}_j$ and the current estimate $\boldsymbol{\Psi}^{(k)}$ for $\boldsymbol{\Psi}$, where

$$\begin{aligned} \tau_{hj}^{(k)} &= Pr_{\boldsymbol{\Psi}^{(k)}}(\zeta_{hj} = 1|y_j, \boldsymbol{w}_j) \\ &= \boldsymbol{\pi}_h^k \frac{f_h^G(\boldsymbol{w}_j; \boldsymbol{\alpha}_h) f_h^E(y_j|\boldsymbol{w}_j; \boldsymbol{\beta}_h)}{\sum_{l=1}^{H} f_l^G(\boldsymbol{w}_j; \boldsymbol{\alpha}_l) f_l^E(y_j|\boldsymbol{w}_j; \boldsymbol{\beta}_l)}, \qquad h = 1 \ldots H. \end{aligned} \tag{3}$$

The Q function can be decomposed into three terms with respect to the parameters $\boldsymbol{\pi}_h$, $\boldsymbol{\alpha}_h$ and

$\boldsymbol{\beta}_h$ to be estimated in the M-step:

$$Q_\pi = \sum_{j=1}^{n}\sum_{h=1}^{H} \tau_{hj}^{(k)} \log \boldsymbol{\pi}_h, \tag{4}$$

$$Q_\alpha = \sum_{j=1}^{n}\sum_{h=1}^{H} \tau_{hj}^{(k)} \log f_h^G(\boldsymbol{w}_j; \boldsymbol{\alpha}_h), \tag{5}$$

$$Q_\beta = \sum_{j=1}^{n}\sum_{h=1}^{H} \tau_{hj}^{(k)} \log f_h^E(y_j|\boldsymbol{w}_j; \boldsymbol{\beta}_h). \tag{6}$$

By maximizing the decomposed Q functions separately in the M-step, one can obtain the updated estimates of $\boldsymbol{\pi}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

1.2 ME networks in practice

The initial estimates of $\boldsymbol{\mu}_h$, $\boldsymbol{\Sigma}_h$ and $\boldsymbol{\pi}_h$ ($h = 1, \ldots, H$) are given by the k-means clustering algorithm on the microarray data, with $k = H$.
The number of experts H can be tuned by computing the index

$$I_h = \sum_{j=1}^{n} \zeta_{hj}/n \simeq \sum_{j=1}^{n} \tau_{hj}/n, \qquad h = 1, \ldots, H.$$

According to Jacobs *et al.* (1997), the number of experts to choose is the minimum value of H for which the sum of the largest indices exceeds 0.8. In practice, in our binary context, we always found that $H = 2$. In fact, many authors already found that the optimal number of experts can often be set to the number of classes (Ubeyli, 2005; Ng and McLachlan, 2007; Gormley *et al.*, 2009).

1.3 Maximization of the Q-function for the ME model

Common unknown parameters for all models. These parameters to estimate are $\boldsymbol{\pi}_h$ and $\boldsymbol{\beta}_h$ in (1), for $h = 1, \ldots, H$.
In the E-step, $\tau_{hj}^{(k)}$ is computed using $f_h^G(\boldsymbol{w}_j; \boldsymbol{\alpha}_h^{(k)})$, which is replaced by (3) using the current estimate $\boldsymbol{\alpha}^{(k)}$.
In the M-step, maximizing Q_π gives:

$$\boldsymbol{\pi}_h^{(k+1)} = \sum_{j=1}^{n} \tau_{hj}^{(k)}/n.$$

The weight vector $\boldsymbol{\beta}_h^{(k+1)}$ in (1) is updated by solving H nonlinear equations using the MINPACK Fortran routine:

$$\sum_{j=1}^{n} \tau_{hj}^{(k)} \left(y_j - \frac{\exp(\boldsymbol{\beta}_h^{(k)T} \boldsymbol{w}_j)}{1 + \exp(\boldsymbol{\beta}_h^{(k)T} \boldsymbol{w}_j)} \right) \boldsymbol{w}_j = 0$$

The parameter $\boldsymbol{\alpha}_h^{(k+1)}$ is estimated by maximizing Q_α for the different types of model that we use, as described below.

Independence model. The vector of unknown parameters $\boldsymbol{\alpha}_h$ consists of $\lambda_{hil}(i = 1, \ldots, q; l = 1, \ldots, n_i - 1)$, and the elements of $\boldsymbol{\mu}_h$ and $\boldsymbol{\Sigma}_h$, $(h = 1, \ldots, H)$.

$$\lambda_{hil}^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{hj}^{(k)} \delta(z_{ij}, l)}{\sum_{j=1}^{n} \tau_{hj}^{(k)}},$$

where $\delta(z_{ij}, l) = 1$ if $z_{ij} = l$ and is zero otherwise, $l = 1, \ldots, n_i$.
The elements $\boldsymbol{\mu}_h$ and $\boldsymbol{\Sigma}_h$ are updated as follows:

$$\boldsymbol{\mu}_h^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{hj}^{(k)} \boldsymbol{x}_j}{\sum_{j=1}^{n} \tau_{hj}^{(k)}},$$

$$\boldsymbol{\Sigma}_h^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{hj}^{(k)} (\boldsymbol{x}_j - \boldsymbol{\mu}_h^{(k)})(\boldsymbol{x}_j - \boldsymbol{\mu}_h^{(k)})^T}{\sum_{j=1}^{n} \tau_{hj}^{(k)}}.$$

Location model. The vector of unknown parameters $\boldsymbol{\alpha}_h$ consists of $p_{hs}(s = 1, \ldots, S)$ and the elements of $\boldsymbol{\mu}_h$ and $\boldsymbol{\Sigma}_h(h = 1, \ldots, H)$. These parameters are estimated as follows:

$$p_{hs}^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{hj}^{(k)} \delta(j, s)}{\sum_{j=1}^{n} \tau_{hj}^{(k)}},$$

where $\delta(j, s) = 1$ if $\boldsymbol{z}_{ij} = s$ and is zero otherwise.The elements $\boldsymbol{\mu}_h$ and $\boldsymbol{\Sigma}_h$ are updated as follows:

$$\boldsymbol{\mu}_h^{(k+1)} = \frac{\sum_{j=1}^{n} \tau_{hj}^{(k)} \delta(j, s) \boldsymbol{x}_j}{\sum_{j=1}^{n} \tau_{hj}^{(k)} \delta(j, s)},$$

$$\boldsymbol{\Sigma}_h^{(k+1)} = \frac{\sum_{j=1}^{n} \sum_{s=1}^{S} \delta(j, s) \tau_{hj}^{(k)} (\boldsymbol{x}_j - \boldsymbol{\mu}_h^{(k)})(\boldsymbol{x}_j - \boldsymbol{\mu}_h^{(k)})^T}{\sum_{j=1}^{n} \tau_{hj}^{(k)}}.$$

Multinomial logit model. The vector of unknown parameters $\boldsymbol{\alpha}_h$ only consists of the variable weight vector $\boldsymbol{v}_h$, which is estimated via the IRLS algorithm outside the EM algorithm, see Jordan and Jacobs (1994).

References

Gormley, I., Murphy, T., *et al.* (2009). A mixture of experts model for rank data with applications in election studies. *Arxiv preprint arXiv:0901.4203*.

Jacobs, R., Peng, F., and Tanner, M. (1997). A Bayesian approach to model selection in hierarchical mixtures-of-experts architectures. *Neural Networks*, **10**(2), 231–241.

Jordan, M. and Jacobs, R. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, **6**(2), 181–214.

Ng, S. and McLachlan, G. (2007). Extension of mixture-of-experts networks for binary classification of hierarchical data. *Artificial Intelligence in Medicine*, **41**(1), 57–67.

Ubeyli, E. (2005). A mixture of experts network structure for breast cancer diagnosis. *Journal of medical systems*, **29**(5), 569–579.