ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Extension of mixture-of-experts networks for binary classification of hierarchical data

Shu-Kay Ng[a,*], Geoffrey J. McLachlan[a,b]

[a] *Department of Mathematics, University of Queensland, Brisbane, Qld 4072, Australia*
[b] *Institute for Molecular Bioscience, University of Queensland, Brisbane, Qld 4072, Australia*

Received 27 January 2007; received in revised form 30 May 2007; accepted 2 June 2007

KEYWORDS
Mixture-of-experts;
Binary classification;
Expectation—maximization algorithm;
Generalized linear mixed-effects model;
Hierarchical data;
Supervised learning

Summary

Objective: For many applied problems in the context of medically relevant artificial intelligence, the data collected exhibit a hierarchical or clustered structure. Ignoring the interdependence between hierarchical data can result in misleading classification. In this paper, we extend the mechanism for mixture-of-experts (ME) networks for binary classification of hierarchical data. Another extension is to quantify cluster-specific information on data hierarchy by random effects via the generalized linear mixed-effects model (GLMM).
Methods and material: The extension of ME networks is implemented by allowing for correlation in the hierarchical data in both the gating and expert networks via the GLMM. The proposed model is illustrated using a real thyroid disease data set. In our study, we consider 7652 thyroid diagnosis records from 1984 to early 1987 with complete information on 20 attribute values. We obtain 10 independent random splits of the data into a training set and a test set in the proportions 85% and 15%. The test sets are used to assess the generalization performance of the proposed model, based on the percentage of misclassifications. For comparison, the results obtained from the ME network with independence assumption are also included.
Results: With the thyroid disease data, the misclassification rate on test sets for the extended ME network is 8.9%, compared to 13.9% for the ME network. In addition, based on model selection methods described in Section 2, a network with two experts is selected. These two expert networks can be considered as modeling two groups of patients with high and low incidence rates. Significant variation among the predicted cluster-specific random effects is detected in the patient group with low incidence rate.
Conclusions: It is shown that the extended ME network outperforms the ME network for binary classification of hierarchical data. With the thyroid disease data, useful

* Corresponding author. Tel.: +61 7 33656139; fax: +61 7 33651477.
E-mail address: skn@maths.uq.edu.au (S.-K. Ng).

doi:10.1016/j.artmed.2007.06.001

information on the relative log odds of patients with diagnosed conditions at different periods can be evaluated. This information can be taken into consideration for the assessment of treatment planning of the disease. The proposed extended ME network thus facilitates a more general approach to incorporate data hierarchy mechanism in network modeling.

1. Introduction

In the context of medically relevant artificial intelligence, many real-world problems involve data that exhibit a hierarchical or clustered structure. These include learning problems in wide areas of biometrical and medical sciences, where a data hierarchy is formed due to the relatedness between multiple tasks. For example, related multiple tasks occur in the prediction of the survival of patients from different hospitals [1]. With these problems, data collected from the same cluster are often interdependent and tend to be more alike in characteristics than data chosen at random from the population as a whole. Ignoring the dependence between hierarchical data can result in overlooking the importance of certain cluster-specific effects and lead to spurious learning or misleading classification [2,3].

Among the various kinds of modular networks, mixtures-of-experts [4] and hierarchical mixtures-of-experts [5] are of much interest due to their wide applicability [6—8] and the advantage of fast learning via the expectation—maximization (EM) algorithm of Dempster et al. [9]; see, for example, [10,11]. The mixture-of-experts (ME) architecture is based on the divide-and-conquer principle where a complex problem is broken up into simpler and smaller problems and their solutions can be combined to yield a solution to the complex problem. Such a strategy can be a powerful tool for modeling mixed tasks with different local rules [5].

In this paper, we extend the mechanism for ME networks for binary classification of hierarchical data via a supervised learning approach. The extension is implemented by allowing both the gating and expert networks with correlations within clusters. Another extension is to quantify cluster-specific information on data hierarchy by random effects via the generalized linear mixed-effects model (GLMM), which is commonly used in the context of statistical multilevel analysis [1,3,12,13]. The predicted cluster-specific random effects provide insights on the comparison between related multiple tasks; for examples, the comparison of the performance of hospitals based on the estimated cluster-specific effects [1]. The proposed model mimics the performance of the human system in that human learning frequently involves approaching several related learning tasks simultaneously and takes advantage of the opportunity to compare and contrast similar multiple tasks in learning for improving generalization accuracy [2]. The remainder of the paper is organized as follows: Section 2 describes the extension of ME networks for binary classification problems with hierarchically structured data via the GLMM. In Section 3, we show how a fast (multitask) learning of the extended ME network can be achieved via the EM algorithm based on a residual maximum likelihood (REML) approach. The proposed model is illustrated in Section 4, using a real thyroid disease data set in the context of binary classification problems. In Section 5, a second example is presented. Related work is discussed in Section 6. Section 7 presents some concluding remarks.

2. Extension of mixture-of-experts for binary classification via GLMM

As shown in Fig. 1(a), the ME architecture is comprised of M expert networks.[1] For binary classification problems, we assume that the output y is a discrete binary indicator variable having possible outcomes of "success" and "failure" [5]. The expert networks approximate the distribution of the output y within each region of the input space. The expert network maps its input $\boldsymbol{x}$ to a local output, the density $f_h(y|\boldsymbol{x};\boldsymbol{\theta}_h)$, where $\boldsymbol{\theta}_h$ is a vector of unknown parameters for the h th expert network $(h = 1, \ldots, M)$. It is assumed that different experts are appropriate in different regions of the input space. The gating network provides a set of scalar coefficients $\pi_h(\boldsymbol{x};\boldsymbol{\alpha})$ that weight the contributions of the various experts, where $\boldsymbol{\alpha}$ is a vector of unknown parameters in the gating network. Therefore, the final output of the ME neural network is a weighted sum of all the local output vectors produced by expert networks:

$$f(y|\boldsymbol{x};\Psi) = \sum_{h=1}^{M} \pi_h(\boldsymbol{x};\boldsymbol{\alpha})\, f_h(y|\boldsymbol{x};\boldsymbol{\theta}_h), \tag{1}$$

[1] This diagram looks somewhat different from the representation in the existing literature so as to contrast with the learning mechanism for extended ME networks in Fig. 1(b).

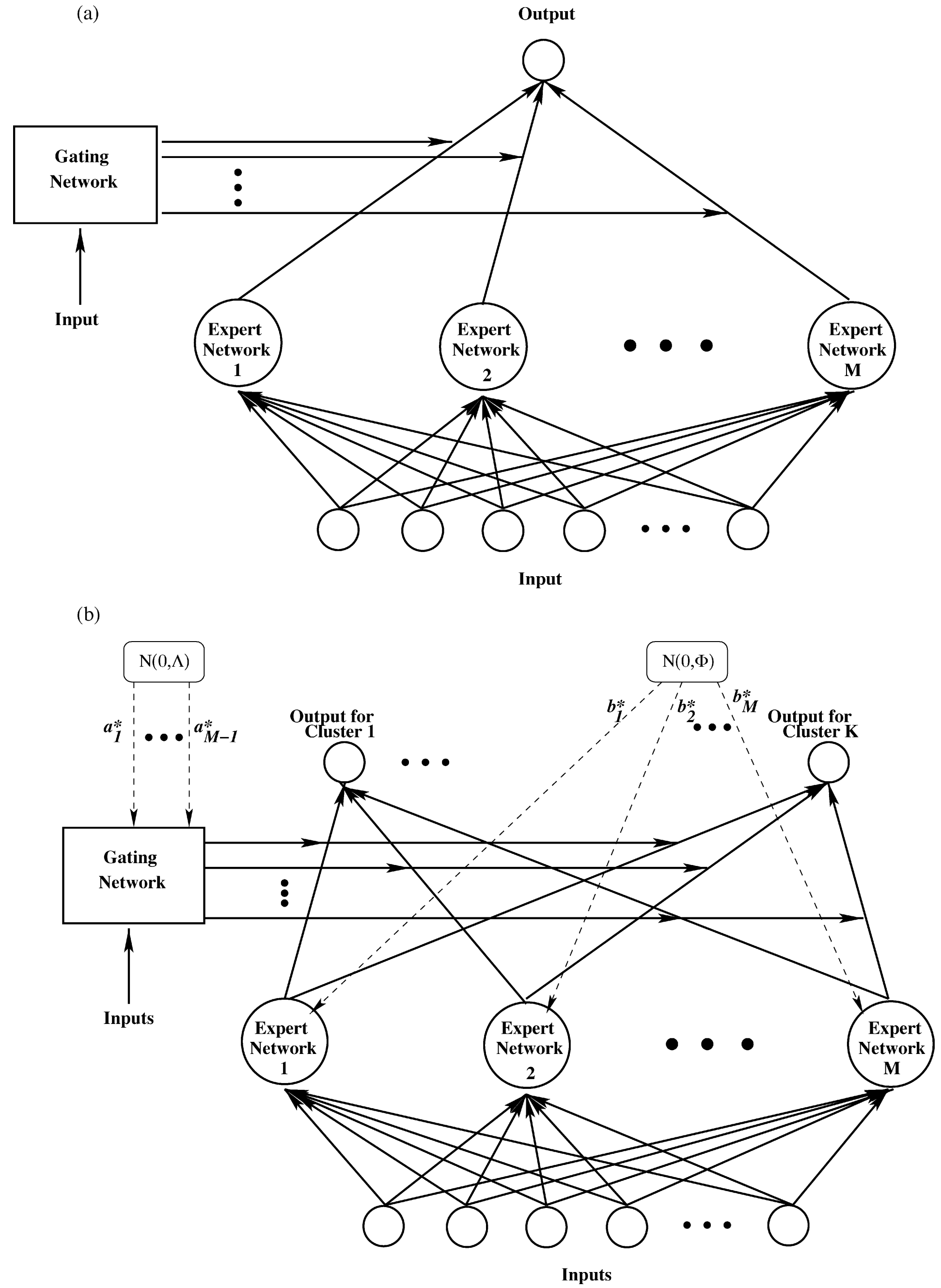

Figure 1 (a) Mixture-of-experts; (b) extended mixture-of-experts with K level-two units. (Here, in contrast to solid lines, dashed lines are used to represent unobservable random effects. The cluster-specific random effects $\boldsymbol{a}_h^*$ and $\boldsymbol{b}_h^*$ are generated from $N(\boldsymbol{0}, \Lambda)$ and $N(\boldsymbol{0}, \Phi)$ for the gating and expert networks, respectively, where $\boldsymbol{a}_h^* = (A_{h1}, \ldots, A_{hK})^T$ and $\boldsymbol{b}_h^* = (B_{h1}, \ldots, B_{hK})^T$.)

where Ψ is the vector of all the unknown parameters [7]. The output of the gating network is modeled by the softmax function as

$$\pi_h(\boldsymbol{x}; \boldsymbol{\alpha}) = \frac{\exp(\boldsymbol{v}_h^T \boldsymbol{x})}{\sum_{l=1}^{M} \exp(\boldsymbol{v}_l^T \boldsymbol{x})} \quad (h = 1, \ldots, M), \qquad (2)$$

where $\boldsymbol{v}_h$ is the weight vector of the h th expert in the gating network and $\boldsymbol{v}_M = \boldsymbol{0}$. It is implicitly assumed that the first element of $\boldsymbol{x}$ is one, to account for an intercept term [5]. In (2), the superscript T denotes vector transpose and $\boldsymbol{\alpha}$ contains the elements in $\boldsymbol{v}_h (h = 1, \ldots, M - 1)$. The unknown parameter vector Ψ can be estimated by the maximum likelihood (ML) approach via the EM algorithm [5,10] or the expectation—conditional maximization (ECM) algorithm [11,14]. With the ME networks, a frequentist analog of the worth index approach [6] may be adopted to select the number of expert networks M; see [7]. The identifiability of ME networks has been studied by Jiang and Tanner [15]. They showed that conditions for identifiability generally hold for some commonly used expert networks such as Poisson, gamma, Gaussian and Bernoulli experts.

For learning problems with hierarchical data structure as mentioned in Section 1, we are given, say for each ith cluster $(i = 1, \ldots, K)$, a data set $D_i = \{\boldsymbol{x}_{i1}, y_{i1}, \ldots, \boldsymbol{x}_{in_i}, y_{in_i}\}$, where n_i is the number of examples for the ith cluster, and K is the total number of clusters. Here, we assume that the input $\boldsymbol{x}_{ij}$ is an p-dimensional vector $(i = 1, \ldots, K;$ $j = 1, \ldots, n_i)$. The complete data set is then given by $D = \{D_i\}$ with the total number of examples $N = \sum_{i=1}^{K} n_i$. In the context of statistical multilevel analysis, this setting corresponds to a two-level hierarchical structure, where the N examples are considered as level-one units and the K clusters as level-two units [3]. The interdependency between hierarchical data from the K clusters (level-two units) can be taken into account by incorporating random effects into the model via the GLMM [12]. That is, we assume that there exists cluster-specific (random) effects, which in turn introduce interdependency among data obtained from the same cluster. This approach has been adopted for the analysis of survival data and regression problems [1,13]. With extended ME networks, we allow both the gating and expert networks to incorporate the data hierarchy via the GLMM (Fig. 1(b)). With the GLMM, the output of the gating network is represented by

$$\pi_h(\boldsymbol{x}_{ij}; \boldsymbol{\alpha}, \boldsymbol{a}, \Lambda) = \frac{\exp(\boldsymbol{v}_h^T \boldsymbol{x}_{ij} + A_{hi})}{1 + \sum_{l=1}^{M-1} \exp(\boldsymbol{v}_l^T \boldsymbol{x}_{ij} + A_{li})} \quad (h = 1, \ldots, M-1), \tag{3}$$

$$\pi_M(\boldsymbol{x}_{ij}; \boldsymbol{\alpha}, \boldsymbol{a}, \Lambda) = \frac{1}{1 + \sum_{l=1}^{M-1} \exp(\boldsymbol{v}_l^T \boldsymbol{x}_{ij} + A_{li})}.$$

In the terminology of GLMM, elements of $\boldsymbol{\alpha}$ (or $\boldsymbol{v}_h$) are fixed effects (unknown constants) that are shared among all clusters, while $A_{hi}(h = 1, \ldots, M-1)$ represent the unobservable cluster-specific random effects from the ith cluster $(i = 1, \ldots, K)$; see, for example, [13]. Letting $\boldsymbol{a} = (\boldsymbol{a}_1^T, \ldots, \boldsymbol{a}_K^T)^T$ and $\boldsymbol{a}_i = (A_{1i}, \ldots, A_{(M-1)i})^T$, we assume that $\boldsymbol{a}$ follows a multivariate normal distribution with zero mean vector and covariance matrix Λ:

$$\Lambda = \begin{bmatrix} \Lambda_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \Lambda_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \cdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \Lambda_K \end{bmatrix}, \tag{4}$$

with $\Lambda_i = \text{diag}(\lambda_1, \ldots, \lambda_{M-1})$ for $i = 1, \ldots, K$, where $\lambda_h(h = 1, \ldots, M-1)$ are known as the variance components in the context of GLMM and assumed to be distinct for different experts.

The inclusion of random effects in expert networks can be handled via the linear predictor in a similar way above. For the h th expert $(h = 1, \ldots, M)$, we let the linear predictor η_{hij} be

$$\eta_{hij} = \boldsymbol{w}_h^T \boldsymbol{x}_{ij} + B_{hi}, \tag{5}$$

where $\boldsymbol{w}_h$ is the weight vector of the h th expert network and $B_{hi}(h = 1, \ldots, M)$ are the unobservable cluster-specific random effects from the ith cluster $(i = 1, \ldots, K)$. Letting $\boldsymbol{b} = (\boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_K^T)^T$ and $\boldsymbol{b}_i = (B_{1i}, \ldots, B_{Mi})^T$, it is assumed that $\boldsymbol{b}$ follows a multivariate normal distribution with zero mean vector and covariance matrix Φ:

$$\Phi = \begin{bmatrix} \Phi_1 & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \Phi_2 & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \cdots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \Phi_K \end{bmatrix}, \tag{6}$$

where $\Phi_i = \text{diag}(\phi_1, \ldots, \phi_M)$ for $i = 1, \ldots, K$, where the variance components $\phi_h(h = 1, \ldots, M)$ are assumed to be different for the M experts. The expected value of the local output is then obtained by passing the linear predictor through the link function $g(\cdot)$ as

$$\boldsymbol{\mu}_{hij} = g(\eta_{hij}). \tag{7}$$

For binary classification problems, the local density $f_h(y_{ij}|\boldsymbol{x}_{ij}; \boldsymbol{\theta}_h, \boldsymbol{b}, \Phi)$ for $i = 1, \ldots, K$ and $j = 1, \ldots, n_i$ is generally assumed to be the Bernoulli distribution and $g(\cdot)$ is the logistic function; see [5]. That is,

$$f_h(y_{ij}|\boldsymbol{x}_{ij}; \boldsymbol{\theta}_h, \boldsymbol{b}, \Phi) = \left(\frac{\exp(\boldsymbol{w}_h^T \boldsymbol{x}_{ij} + B_{hi})}{1 + \exp(\boldsymbol{w}_h^T \boldsymbol{x}_{ij} + B_{hi})} \right)^{y_{ij}} \times \left(\frac{1}{1 + \exp(\boldsymbol{w}_h^T \boldsymbol{x}_{ij} + B_{hi})} \right)^{1-y_{ij}}, \tag{8}$$

where the vector of unknown parameters for the h th expert network is equal to the weight vector $(\boldsymbol{\theta}_h = \boldsymbol{w}_h)$. These fixed effects are shared among all cluster (Fig. 1(b)).

An advantage of the use of the GLMM is that the predicted cluster-specific random effects have a meaningful interpretation. For example, the estimates of random effects $\boldsymbol{a}_i(i = 1, \ldots, K)$ in the gating network (3) quantify the extent of cluster-specific effect from the ith cluster on the weights of the contributions from various experts. Similarly, the estimates of random effects $\boldsymbol{b}_i(i = 1, \ldots, K)$ in expert networks (5) provide useful information as to whether there is a significant difference in local outputs from each expert network among data from different clusters. These quantified cluster-specific information can be used to draw insights on the comparison between different clusters; see, for example, [1]. We shall illustrate this issue further in Section 4.

3. Learning via the ECM algorithm

With the extended ME network for hierarchical data, we let $\Psi^T = (\Psi_1^T, \Psi_2^T)$ be the vector of unknown parameters, where $\Psi_1^T = (\boldsymbol{\alpha}^T, \boldsymbol{a}^T, \boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_M^T, \boldsymbol{b}^T)$ consists of the fixed and the unobservable random effects, and Ψ_2 contains the variance components $\lambda_h (h = 1, \ldots, M-1)$ and $\phi_h (h = 1, \ldots, M)$ in Λ and Φ, respectively, for the gating and expert networks. The supervised learning of the unknown parameters Ψ within the extended ME network is implemented via the REML approach of McGilchrist [12] and its extension for normal mixture models [13]. For given initial values of Λ and Φ, the best linear unbiased prediction (BLUP) estimators of Ψ_1 maximize the function[2] $L = L_1 + L_2$, where

L_1 = log likelihood formed from output $\boldsymbol{Y}_{ij}$ with $\boldsymbol{a}$ and $\boldsymbol{b}$ conditionally fixed,

L_2 = logarithm of the joint probability density *function of* $\boldsymbol{a}$ and $\boldsymbol{b}$, with $\boldsymbol{a}$ and $\boldsymbol{b}$ taken to be independent. (9)

The BLUP estimators are then used to obtain approximate REML estimators of the parameters Ψ_2 for the variance components [12,16]. The BLUP estimate of Ψ_1 is obtained as a solution of the equation $\partial L/\partial\Psi_1 = 0$, which can be solved via the ECM algorithm [11,14,17]. The ECM algorithm is a broadly applicable technique that provides an iterative procedure for computing ML estimates in a variety of incomplete-data problems such as the learning of ME networks [11].

In order to pose the learning for the extended ME network as an incomplete-data problem, we introduce the indicator variables z_{hij}, where z_{hij} is one or zero according to whether $\boldsymbol{y}_{ij}$ belongs or does not belong to the h th expert $(h = 1, \ldots, M)$. For the joint log likelihood $L = L_1 + L_2$ in (9), the BLUP estimate of Ψ_1 can be found iteratively using the ECM algorithm as detailed in [11]. Given the current estimates of $\Psi_1^{(c)}$ and $\Psi_2^{(c)}$, the expectation (E) step on the $(c+1)$ th iteration involves the computation of the Q-function which is given by the expected value of the complete-data log likelihood conditional on the observed data and the current model. That is, with reference to (8):

$$\begin{aligned}
Q(\Psi_1; \Psi_1^{(c)}, \Psi_2^{(c)}) &= \sum_{i=1}^{K}\sum_{j=1}^{n_i}\sum_{h=1}^{M} \tau_{hij}^{(c)}\{\log \pi_h(\boldsymbol{x}_{ij}; \boldsymbol{\alpha}, \boldsymbol{a}, \Lambda^{(c)}) \\
&\quad + \boldsymbol{y}_{ij}(\boldsymbol{w}_h^T \boldsymbol{x}_{ij} + \boldsymbol{B}_{hi}^{(c)}) \\
&\quad - \log(\boldsymbol{1} + \exp(\boldsymbol{w}_h^T \boldsymbol{x}_{ij} + \boldsymbol{B}_{hi}^{(c)}))\} \\
&\quad - \frac{1}{2}\{\log|\Lambda^{(c)}| + \boldsymbol{a}^T \Lambda^{(c)^{-1}} \boldsymbol{a}\} \\
&\quad - \frac{1}{2}\{\log|\Phi^{(c)}| + \boldsymbol{b}^T \Phi^{(c)^{-1}} \boldsymbol{b}\} - C,
\end{aligned} \tag{10}$$

where $C = K\log(2\pi)$ is a constant and

$$\begin{aligned}
\tau_{hij}^{(c)} &= \mathrm{pr}\{Z_{hij} = 1 | y_{ij}, \boldsymbol{x}_{ij}, \Psi_1^{(c)}, \Psi_2^{(c)}\} \\
&= \frac{\pi_h(\boldsymbol{x}_{ij}; \boldsymbol{\alpha}^{(c)}, \boldsymbol{a}^{(c)}, \Lambda^{(c)})(((\exp(\boldsymbol{w}_h^{(c)^T} \boldsymbol{x}_{ij} + \boldsymbol{B}_{hi}^{(c)}))^{y_{ij}})/(\boldsymbol{1} + \exp(\boldsymbol{w}_h^{(c)^T} \boldsymbol{x}_{ij} + \boldsymbol{B}_{hi}^{(c)})))}{\sum_{l=1}^{M} \pi_l(\boldsymbol{x}_{ij}; \boldsymbol{\alpha}^{(c)}, \boldsymbol{a}^{(c)}, \Lambda^{(c)})(((\exp(\boldsymbol{w}_l^{(c)^T} \boldsymbol{x}_{ij} + \boldsymbol{B}_{li}^{(c)}))^{y_{ij}})/(\boldsymbol{1} + \exp(\boldsymbol{w}_l^{(c)^T} \boldsymbol{x}_{ij} + \boldsymbol{B}_{li}^{(c)})))}
\end{aligned} \tag{11}$$

is the current estimated posterior probability that y_{ij} belongs to the h th expert $(h = 1, \ldots, M)$. For Bernoulli distributions (8), it can be seen from (10) that the Q-function can be gathered into a term with respect to the gating network and M terms corresponding for each expert of the expert network; see the discussion in [11].

The updated estimates $\Psi_1^{(c+1)}$ are obtained in the M-step by maximizing the Q-function (10) over the parameter space. As the Q-function can be decomposed into separate terms corresponding to the gating and each expert network, it implies that separate maximizations can be performed independently and fast learning can be achieved. The detailed description of the M-step for Bernoulli models is presented in Appendix A.

Given the updated BLUP estimates $\Psi_1^{(c+1)}$, the approximate REML estimates $\Psi_2^{(c+1)}$ of the variance components Λ and Φ are obtained based on the procedure described in [12,16]. Asymptotic variances of the estimators $\hat{\Psi}_2$ of the variance components can be obtained from the inverse of the REML information matrix [16], which are used to

[2] The function L is not a log likelihood in the conventional sense because it is based on the unobservable random effects $\boldsymbol{a}$ and $\boldsymbol{b}$.

assess if clusters differ significantly in the weighting of experts and the mean of local outputs. Appendix B outlines the REML procedure for Ψ_2 with the extended ME network.

4. An example of thyroid disease data

In this section, the extended ME network is applied to a real thyroid disease data set. The data set "thyroid0387.data" is available from the UCI Repository of machine learning databases [18], consisting of 9172 thyroid diagnosis records from 1984 to early 1987. Each record has 29 attribute values and a thyroid diagnosis. The diagnosis covers 20 classes, but here we consider a binary outcome variable that indicates the presence or the absence of diagnosed conditions. There are plenty of missing values (5.3%) in the original data set. In our study, we consider 7652 records with complete information on 20 attribute values (15 binary variables and five continuous variables). Based on the record identification number, we create a hierarchical data structure with $K = 26$ level-two units. These level-two units (clusters) thus represent diagnosis records from different periods between 1984 and early 1987. The values of each continuous attribute are scaled to have zero mean and unit variance.

Based on the model selection method described in [7], the number of experts is chosen to be the minimum number of experts with the largest worth indices for which the sum of their worth indices exceeds some critical value κ, says, $\kappa = 0.8$ [6]. With the ME networks, result of applying this model selection method to the thyroid disease data is presented in Table 1. For comparison, we include also the model selection approach based on the Bayesian information criterion (BIC) [19]. Based on the result presented in Table 1, a ME network with $M = 2$ is selected.

For the study of the applicability of the proposed model, we obtain ten independent random splits of the data into a training set and a test set, in a proportion of 85% and 15%, respectively. The test sets are used to assess the generalization performance of the proposed model, based on the percentage of misclassifications by the model. The results are presented in Table 2. It can be seen that the extended ME network provides a smaller averaged number of misclassified data and provides better performance in the binary classification of this thyroid disease data.

Table 2 Classification results for the thyroid disease data

Model	No. misclassified on test sets[a]	Percentage of misclassification on test sets[a]
ME network	159±14	13.9±1.3
Extended ME network	102±15	8.9±1.3

[a] Mean ± standard deviation.

As described at the end of Section 2, the predicted cluster-specific random effects ***a*** and ***b*** have a meaningful interpretation. With the thyroid disease data, the two expert networks can be considered as modeling two groups of patients with high and low probabilities of the presence of diagnosed conditions (incidence rates), respectively. Based on the asymptotic variances of the estimators in the variance components obtained according to the procedure described in Appendix B, significance of the cluster effects can be assessed. With the thyroid data, significant variations among the predicted cluster-specific random effects in the patient group with low incidence rate is detected. In Fig. 2, we display the predicted cluster-specific random effects B_{1i} and B_{2i} (significant variation) in expert networks for these two patient groups among the 26 clusters. As a result, an estimated negative random effect B_{2i} for the patient group with low incidence rate thus indicates a smaller log odds of the presence of diagnosed conditions in a cluster, under a Bernoulli probability model (8). Based on the predicted cluster-specific random effects, the relative log odds of patients with diagnosed conditions at different period can be evaluated. This useful information can be taken into consideration for the assessment of treatment planning of the disease.

5. An example on multi-center clinical trials

To illustrate further the applicability of the proposed model, a second example is presented. The

Table 1 Model selection for the thyroid disease data

No. of experts	log likelihood	BIC	Worth indices
2	−2711.87	5987.1[a]	(0.624, 0.376)[a]
3	−2607.04	6153.1	(0.50, 0.33, 0.17)
4	−2521.33	6357.2	(0.47, 0.44, 0.05, 0.04)

[a] The number of experts selected by each model selection method.

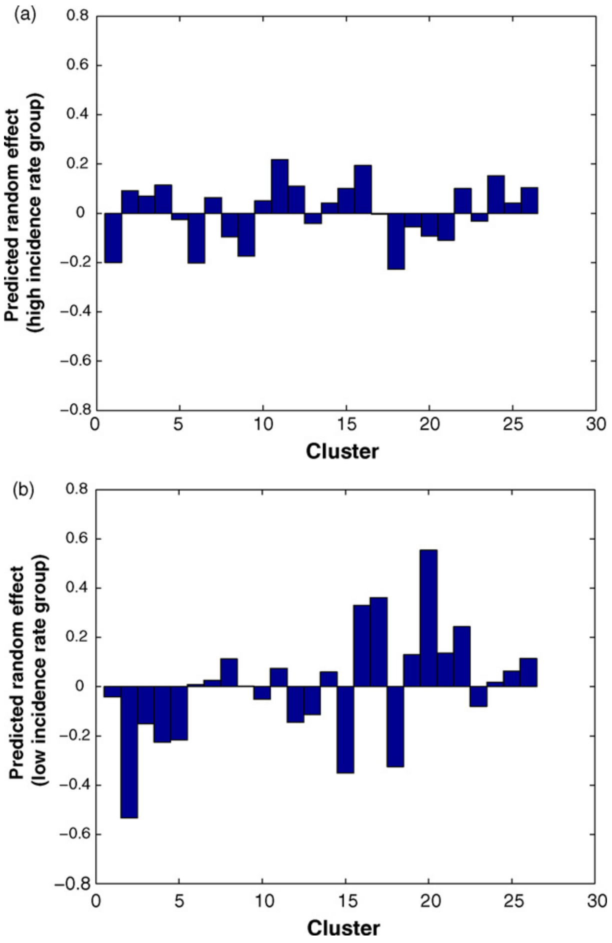

Figure 2 Prediction of cluster-specific random effects for (a) the patient group with high incidence rate and (b) the patient group with low incidence rate.

data set is available in [20], which is a part of a large multi-center clinical trial carried out by the Radiation Therapy Oncology Group in the United States. The data set in [20] involved the patients with squamous carcinoma of three sites in the oropharynx, with six institutions participating. Each treatment policy dictated the treatment to be administrated during a 90-day period. After this period, each patient received medical care by the participating institution [20]. As there was considerable variability in patient treatment following the 90-day period and in the facilities shared within participating institution, it is conceived that institution effect may also be of importance in the analysis of this multi-center clinical trial data. In our study, we consider only the carcinoma of the pharyngeal tongue ($N = 59$) with complete information on four attribute values (three binary variables and one continuous variable). A hierarchical data structure with $K = 6$ level-two units, corresponding to the six participating institutions, is postulated. We consider here a binary outcome variable that indicates a patient's survival time being greater than three years or not. Censored observations corresponding to patients that were lost to follow-up within three years are ignored. The result for the model selection is given in Table 3(a). Based on Table 3(a), a ME network with $M = 2$ is selected.

The performance of the proposed model is assessed based on the "leave-one-out" misclassification error rate for an "external" cross-validation [21]. The leave-one-out procedure can be viewed as the special case where the size of the test set is reduced to a single entity. As the parameters are trained based on the training set after eliminating a single entity each time, this external leave-one-out procedure thus reduces the bias in estimating the misclassification error rate [21]. The results are presented in Table 3(b). It can be seen that the extended ME network provides a smaller leave-one-out misclassification error rate, compared to the ME network. Based on the asymptotic variances of the estimators in the variance components, the institution effect is not significant.

Table 3 The multi-center clinical trial data

(a) Model selection

No. of experts	log likelihood	BIC	Worth indices
2	−26.66	126.7[a]	(0.565, 0.435)[a]
3	−22.85	168.0	(0.50, 0.41, 0.09)
4	−21.92	215.1	(0.44, 0.26, 0.18, 0.12)

(b) Leave-one-out misclassification rate

Model	Number (percentage) misclassified
ME network	10 (16.9%)
Extended ME network	7 (11.9%)

[a] The number of experts selected by each model selection method.

6. Discussion

The advantage of the learning mechanism for extended ME networks relies on the relatedness between multiple tasks. Experimental work [2] has validated this mechanism with sets of subtasks related in various ways. In this paper, we focus on related multiple tasks in binary classification problems which are arisen from a setting of the multilevel analysis of hierarchical data [3]. That is, multiple data sets corresponding to different clusters are obtained and it is anticipated that data from the same cluster are interdependent and tend to be more alike in characteristics than data chosen at random from the population as a whole. The dependency within clusters is incorporated into the network modeling via the GLMM.

The extension of ME networks via the GLMM can be related to the heterogeneity model or the latent class model in the statistical literature [22,23]. For the estimation of variance components in Ψ_2, various approaches have been proposed, including Bayesian analysis [22,24] and the REML estimation procedure of McGilchrist [12], among others. In this paper, the adoption of the REML approach facilitates the estimation of the variance components. It has been shown in the literature that the REML estimation of the variance components provides less biased estimators compared to the ML method, including successful applications in the analysis of survival data and regression problems [1,12,13,16]. With binary classification problems focussed in this paper, the BLUP of Ψ_1 cannot be updated independently. The supervised learning procedure for the estimate of Ψ_1 is therefore performed in a conditional mode using an iterative reweighted least squares (IRLS) approach within the ECM algorithm framework (Appendix A). This is in contrast to the unsupervised learning for regression problems described in [13,16].

The idea of incorporating random effects in multilayered perceptron networks has been considered in [2] in the context of multitask learning. Unlike our method, they incorporate the random effects in the weight vectors of the hidden units. Denoting n_{hidden} the number of hidden units in the neural network model, each random effect, a $(n_{\text{hidden}} + 1)$-dimensional vector, represents the hidden-to-output weights for each subtask. In contrast to the GLMM approach proposed in this paper, the incorporation of random effects in the weight vectors does not possess a meaningful interpretation. The GLMM is a natural extension of the generalized linear model (GLM), in the specifications of the gating and experts networks [5], to incorporate random effects via the corresponding linear predictors. As described in Sections 2 and 3, the GLMM provides a statistically principled approach to quantify the extent of influence from each cluster on both the gating and experts networks via a "soft sharing" mechanism. These predicted random effects are useful on the comparison between different clusters for decision making.

7. Conclusions

Many real-world problems in wide areas of medically relevant artificial neural network applications involve data that exhibit a hierarchical or clustered structure. In this paper, we have extended the supervised learning mechanism for ME networks to tackle binary classification problems with hierarchically structured data. The cluster-specific effects are assumed to be random and modeled via linear predictors, based on the GLMM. This approach provides an alternative method in incorporating random effects within network modeling. For example, the method of incorporating random effects via the GLMM can be applied to learn other alternative models with the ME architecture, such as the mixture of Cox experts [8] and the normalized Gaussian ME model [25]. The former model combines features of the Cox proportional hazards model and the ME networks for modeling survival data with censored observations, while the latter model has a normalized Gaussian gating network. As described in Section 1, the extended ME network would have wide application in various scientific fields where binary classification of hierarchical data is involved.

In Sections 4 and 5, the proposed extended ME network is illustrated using real examples of thyroid disease data and multi-center clinical trial data. It is shown that significant improvement in the misclassification rate is achieved by the adoption of the extended ME network to tackle problems with hierarchical structured data where interdependence of data collected from the same cluster exists. In addition, the proposed GLMM for binary classification problems provides a meaningful interpretation of the predicted cluster-specific random effects for decision making. With the thyroid disease data, significant variation in the cluster-specific random effects is detected in the patient group with low incidence rate. Useful information on the relative log odds of patients with diagnosed conditions at different period can be evaluated based on the predicted cluster-specific random effects. This information can be taken into consideration for the assessment of treatment planning of the disease.

Acknowledgements

The authors would like to acknowledge support for this research by a grant from the Australian Research Council. Thanks are also due to the editor and three anonymous reviewers for their helpful comments and suggestions.

Appendix A. Maximization of the Q-function for Bernoulli models

In this appendix, we describe how the Q-function (10) of the extended ME model with Bernoulli local densities (8) can be maximized on the M-step via the ECM algorithm [11]. With the ECM algorithm and the independence assumption of Λ in (4), we partition the parameter vector $\boldsymbol{\alpha}$ as $(\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_{M-1}^T)^T$, where each $\boldsymbol{\alpha}_h$ consists of the weight vector $\boldsymbol{v}_h$ and the random effect vector $\boldsymbol{a}_h^* = (A_{h1}, \ldots, A_{hK})^T$ of the h th expert. That is, $\boldsymbol{\alpha}_h^T = (\boldsymbol{v}_h^T, \boldsymbol{a}_h^{*^T})$ for $h = 1, \ldots, M-1$, where $\boldsymbol{a} = (\boldsymbol{a}_1^{*^T}, \ldots, \boldsymbol{a}_{(M-1)}^{*^T})^T$ is regrouped according to the order of experts from $h = 1$ to $h = M-1$. On the $(c+1)$ th iteration of the ECM algorithm, the M-step is replaced by $(M-1)$ computationally simpler conditional-maximization (CM) steps:

- **CM-step 1:** Calculate $\boldsymbol{\alpha}_1^{(c+1)}$ by maximizing Q_α with $\boldsymbol{\alpha}_l (l = 2, \ldots, M-1)$ fixed at $\boldsymbol{\alpha}_l^{(c)}$,
- **CM-step 2:** Calculate $\boldsymbol{\alpha}_2^{(c+1)}$ by maximizing Q_α with $\boldsymbol{\alpha}_1$ fixed at $\boldsymbol{\alpha}_1^{(c+1)}$ and $\boldsymbol{\alpha}_l (l = 3, \ldots, M-1)$ fixed at $\boldsymbol{\alpha}_l^{(c)}$,
- $\vdots$
- **CM-step$(M-1)$:** Calculate $\boldsymbol{\alpha}_{(M-1)}^{(c+1)}$ by maximizing Q_α with $\boldsymbol{\alpha}_l (l = 1, \ldots, M-2)$ fixed at $\boldsymbol{\alpha}_l^{(c+1)}$,

where

$$Q_\alpha = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \sum_{h=1}^{M} \tau_{hij}^{(c)} \log \pi_h(\boldsymbol{x}_{ij}; \boldsymbol{\alpha}, \boldsymbol{a}, \Lambda^{(c)}) - \frac{1}{2}\{K \log(2\pi) + \log|\Lambda^{(c)}| + \boldsymbol{a}^T \Lambda^{(c)^{-1}} \boldsymbol{a}\}$$

is the term of the Q-function in (10) for the gating network. The ECM algorithm preserves the appealing convergence properties of the EM algorithm, such as the monotone increasing of likelihood after each iteration [26]; see also [11,14]. More importantly, each CM-step above corresponds to a separable set of the parameters in $\boldsymbol{\alpha}_h$ for $h = 1, \ldots, M-1$, and can be obtained using an iterative reweighted least squares (IRLS) approach [5].

Let $\boldsymbol{\alpha}^{(c+h/(M-1))} = (\boldsymbol{\alpha}_1^{(c+1)^T}, \ldots, \boldsymbol{\alpha}_{h-1}^{(c+1)^T}, \boldsymbol{\alpha}_h^{(c)^T}, \ldots, \boldsymbol{\alpha}_{M-1}^{(c)^T})^T$, at the h th CM-step on the $(c+1)$ th iteration of the ECM algorithm $(h = 1, \ldots, M-1)$, it follows from (3) and (4) that the IRLS updating rule for $\boldsymbol{\alpha}_h$ is given by

$$\boldsymbol{\alpha}_h^{(c+1)} = \boldsymbol{\alpha}_h^{(c)} + \left[-\frac{\partial^2 Q_\alpha}{\partial \boldsymbol{\alpha}_h \boldsymbol{\alpha}_h^T}\right]^{-1}_{(c+h/(M-1))} \left[\frac{\partial Q_\alpha}{\partial \boldsymbol{\alpha}_h}\right]_{(c+h/(M-1))}. \qquad (12)$$

Letting $\boldsymbol{X}$ and $\boldsymbol{S}$ denote the design matrices of $\boldsymbol{v}_h$ and $\boldsymbol{a}_h^*$, respectively, we have

$$\left[\frac{\partial Q_\alpha}{\partial \boldsymbol{\alpha}_h}\right]_{(c+h/(M-1))} = \begin{bmatrix} \boldsymbol{X}^T \\ \boldsymbol{S}^T \end{bmatrix} \boldsymbol{G} - \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{a}_h^{*^{(c)}} / \lambda_h^{(c)} \end{bmatrix}$$

$$\left[-\frac{\partial^2 Q_\alpha}{\partial \boldsymbol{\alpha}_h \boldsymbol{\alpha}_h^T}\right]_{(c+h/(M-1))} = \begin{bmatrix} \boldsymbol{X}^T \\ \boldsymbol{S}^T \end{bmatrix} \boldsymbol{U}_\alpha [\boldsymbol{X} \quad \boldsymbol{S}] + \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_K / \lambda_h^{(c)} \end{bmatrix},$$

where $\boldsymbol{I}_K$ is an identity matrix of dimension K and $\boldsymbol{G}$ is a N by 1 matrix with elements

$$\tau_{hij}^{(c)} - \pi_{hij}^{(c+h/(M-1))} \quad (i = 1, \ldots, K;\ j = 1, \ldots, n_i),$$

and where $\pi_{hij}^{(c+h/(M-1))}$ denotes $\pi_h(\boldsymbol{x}_{ij}; \boldsymbol{\alpha}^{(c+h/(M-1))}, \boldsymbol{a}, \Lambda^{(c)})$. The matrix $\boldsymbol{U}_\alpha$ is a $N \times N$ diagonal matrix with diagonal elements:

$$\pi_{hij}^{(c+h/(M-1))}(1 - \pi_{hij}^{(c+h/(M-1))}) \quad (i = 1, \ldots, K;\ j = 1, \ldots, n_i).$$

The IRLS loop (12) is referred to as the inner loop of the EM algorithm [5]. It is terminated when the algorithm has converged or after some prespecified number of iterations, say, ten iterations.

In the applications to binary classification problems, a Bernoulli model with a logistic link function is used. Let the term of the Q-function in (10) for expert networks be

$$Q_\theta = \sum_{i=1}^{K} \sum_{j=1}^{n_i} \sum_{h=1}^{M} \tau_{hij}^{(c)} \log\ f_h(y_{ij} | \boldsymbol{x}_{ij}; \boldsymbol{\theta}_h, \boldsymbol{b}, \Phi^{(c)}) - \frac{1}{2}\{K \log(2\pi) + \log|\Phi^{(c)}| + \boldsymbol{b}^T \Phi^{(c)^{-1}} \boldsymbol{b}\},$$

where $f_h(y_{ij} | \boldsymbol{x}_{ij}; \boldsymbol{\theta}_h, \boldsymbol{b}, \Phi^{(c)})$ is given by (8). With the independence assumption of Φ in (6), we let $\boldsymbol{b}_h^* = (B_{h1}, \ldots, B_{hK})^T$ contain the random effects for the h th expert, where $\boldsymbol{b} = (\boldsymbol{b}_1^{*^T}, \ldots, \boldsymbol{b}_M^{*^T})^T$ is regrouped according to the order of experts from $h = 1$ to $h = M$. From (6) and (10), we have

$$\begin{bmatrix} \boldsymbol{w}_h^{(c+1)} \\ \boldsymbol{b}_h^{*^{(c+1)}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}_h^{(c)} \\ \boldsymbol{b}_h^{*^{(c)}} \end{bmatrix} + \left[-\frac{\partial^2 Q_\theta}{\partial \boldsymbol{\theta}_h \boldsymbol{\theta}_h^T}\right]^{-1} \left[\frac{\partial Q_\theta}{\partial \boldsymbol{\theta}_h}\right], \qquad (13)$$

where

$$\left[\frac{\partial Q_\theta}{\partial \boldsymbol{\theta}_h}\right] = \begin{bmatrix} \boldsymbol{X}^T \\ \boldsymbol{S}^T \end{bmatrix} \boldsymbol{G} - \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{b}_h^{*^{(c)}} / \phi_h^{(c)} \end{bmatrix}$$

and

$$\left[-\frac{\partial^2 Q_\theta}{\partial \boldsymbol{\theta}_h \boldsymbol{\theta}_h^T}\right] = \begin{bmatrix} \boldsymbol{X}^T \\ \boldsymbol{S}^T \end{bmatrix} \boldsymbol{U}_\theta [\boldsymbol{X} \quad \boldsymbol{S}] + \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & I_K/\phi_h^{(c)} \end{bmatrix}.$$

Here, $\boldsymbol{G}$ is a N by 1 matrix with elements:

$$\tau_{hij}^{(c)} - \frac{\exp \eta_{hij}^{(c)}}{1+\exp \eta_{hij}^{(c)}} \quad (i=1,\ldots,K;\; j=1,\ldots,n_i),$$

where

$$\eta_{hij}^{(c)} = \boldsymbol{w}_h^{(c)^T} \boldsymbol{x}_{ij} + B_{hi}^{(c)}.$$

The matrix $\boldsymbol{U}_\theta$ is a $N \times N$ diagonal matrix with diagonal elements:

$$\tau_{hij}^{(c)} \left(\frac{\exp \eta_{hij}^{(c)}}{1+\exp \eta_{hij}^{(c)}}\right)\left(\frac{1}{1+\exp \eta_{hij}^{(c)}}\right)$$

for $i=1,\ldots,K$ and $j=1,\ldots,n_i$.

Appendix B. REML estimation of variance components

Given the updated BLUP estimates $\Psi_1^{(c+1)}$, the approximate REML estimates of the variance components Λ and Φ are obtained based on the procedure described in [12] and [16]. With the independence assumptions of Λ and Φ in (4) and (6), respectively, we let Π denote the negative second derivative of $L = L_1 + L_2$ in (9) with respect to $\boldsymbol{v}|\boldsymbol{w}|\boldsymbol{a}|\boldsymbol{b}$ in the BLUP procedure (Section 3), where $\boldsymbol{v}^T = (\boldsymbol{v}_1^T, \ldots, \boldsymbol{v}_{M-1}^T)$, $\boldsymbol{w}^T = (\boldsymbol{w}_1^T, \ldots, \boldsymbol{w}_M^T)$, $\boldsymbol{a}^T = (\boldsymbol{a}_1^{*^T}, \ldots, \boldsymbol{a}_{(M-1)}^{*^T})$, and $\boldsymbol{b}^T = (\boldsymbol{b}_1^{*^T}, \ldots, \boldsymbol{b}_M^{*^T})$. Letting $\Pi^{-1} = \boldsymbol{H}$ where the matrix $\boldsymbol{H}$ is partitioned conformally to $\boldsymbol{v}|\boldsymbol{w}|\boldsymbol{a}|\boldsymbol{b}$, it follows from [16] that the elements of Λ and Φ are, respectively, given by

$$\hat{\lambda}_h = K^{-1}(tr\boldsymbol{H}_{ah} + \hat{\boldsymbol{a}}_h^{*^T} \hat{\boldsymbol{a}}_h^*) \quad (h=1,\ldots,M-1)$$

and

$$\hat{\phi}_h = K^{-1}(tr\boldsymbol{H}_{bh} + \hat{\boldsymbol{b}}_h^{*^T} \hat{\boldsymbol{b}}_h^*) \quad (h=1,\ldots,M),$$

where $\boldsymbol{H}_{ah}$ is the $K \times K$ matrix corresponding to the h th partition $(h=1,\ldots,M-1)$ of the part of the original matrix $\boldsymbol{H}$ partitioned conformally with respect to $\boldsymbol{a}$. The $K \times K$ matrix $\boldsymbol{H}_{bh}$ corresponds to the h th partition $(h=1,\ldots,M)$ of the part of $\boldsymbol{H}$ partitioned conformally with respect to $\boldsymbol{b}$.

Asymptotic variances of the estimators $\hat{\Psi}_2$ in the variance components Λ and Φ can be obtained from the inverse of the REML information matrix [16]. For example, with a network of $M = 2$experts, the part of $\boldsymbol{H}$ partitioned conformally with respect to $\boldsymbol{a}$ and $\boldsymbol{b}$ is

$$\begin{bmatrix} \boldsymbol{H}_{a1} & \boldsymbol{H}_{ab1} & \boldsymbol{H}_{ab2} \\ \cdot & \boldsymbol{H}_{b1} & \boldsymbol{H}_{bb1} \\ \cdot & \cdot & \boldsymbol{H}_{b2} \end{bmatrix}.$$

The asymptotic covariance matrix for the variance components $\hat{\lambda}_1$, $\hat{\phi}_1$, and $\hat{\phi}_2$ is then given by

$$\operatorname{cov}\begin{bmatrix} \hat{\lambda}_1 \\ \hat{\phi}_1 \\ \hat{\phi}_2 \end{bmatrix} = 2\begin{bmatrix} \boldsymbol{R}_{a1} & \boldsymbol{R}_{ab1} & \boldsymbol{R}_{ab2} \\ \cdot & \boldsymbol{R}_{b1} & \boldsymbol{R}_{bb1} \\ \cdot & \cdot & \boldsymbol{R}_{b2} \end{bmatrix}^{-1},$$

where the diagonal elements are

$$\boldsymbol{R}_{a1} = \hat{\lambda}_1^{-2}\left\{tr\left(I_K - \frac{\boldsymbol{H}_{a1}}{\hat{\lambda}_1}\right)^2\right\}$$

and

$$\boldsymbol{R}_{bh} = \hat{\phi}_h^{-2}\left\{tr\left(I_K - \frac{\boldsymbol{H}_{bh}}{\hat{\phi}_h}\right)^2\right\} \quad (h=1,2).$$

The off-diagonal elements are given by

$$\boldsymbol{R}_{abh} = \hat{\lambda}_1^{-1}\hat{\phi}_h^{-1} tr(\boldsymbol{H}_{abh}\boldsymbol{H}_{abh}^T) \quad (h=1,2)$$

and

$$\boldsymbol{R}_{bb1} = \hat{\phi}_1^{-1}\hat{\phi}_2^{-1} tr(\boldsymbol{H}_{bb1}\boldsymbol{H}_{bb1}^T).$$

References

[1] Ng S-K, McLachlan GJ, Yau KKW, Lee AH. Modelling the distribution of ischaemic stroke-specific survival time using an EM-based mixture approach with random effects adjustment. Stat Med 2004;23:2729—44.

[2] Bakker B, Heskes T. Task clustering and gating for Bayesian multitask learning. J Mach Learn Res 2003;4:83—99.

[3] Goldstein H. Multilevel statistical models, 2nd ed, London: Arnold; 1995.

[4] Jacobs RA, Jordan MI, Nowlan SJ, Hinton GE. Adaptive mixtures of local experts. Neural Comput 1991;3:79—87.

[5] Jordan MI, Jacobs RA. Hierarchical mixtures of experts and the EM algorithm. Neural Comput 1994;6:181—214.

[6] Jacobs RA, Peng F, Tanner MA. A Bayesian approach to model selection in hierarchical mixtures-of-experts architectures. Neural Networks 1997;10:231—41.

[7] Ng S-K, McLachlan GJ, Lee AH. An incremental EM-based learning approach for on-line prediction of hospital resource utilization. Artif Intell Med 2006;36:257—67.

[8] Rosen O, Tanner MA. Mixtures of proportional hazards regression models. Stat Med 1999;18:1119—31.

[9] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). J Roy Stat Soc B 1977;39:1—38.

[10] Jordan MI, Xu L. Convergence results for the EM approach to mixtures of experts architectures. Neural Networks 1995;8:1409—31.

[11] Ng S-K, McLachlan GJ. Using the EM algorithm to train neural networks: misconceptions and a new algorithm for

multiclass classification. IEEE Trans Neural Networks 2004;15:738—49.

[12] McGilchrist CA. Estimation in generalized mixed models. J Roy Stat Soc B 1994;56:61—9.

[13] Yau KKW, Lee AH, Ng S-K. Finite mixture regression model with random effects: application to neonatal hospital length of stay. Comput Stat Data Anal 2003;41:359—66.

[14] Meng XL, Rubin DB. Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika 1993;80: 267—78.

[15] Jiang W, Tanner MA. On the identifiability of mixtures-of-experts. Neural Networks 1999;12:1253—8.

[16] McGilchrist CA, Yau KKW. The derivation of BLUP, ML, REML estimation methods for generalized linear mixed models. Commun Stat-Theor Meth 1995;24:2963—80.

[17] Ng S-K, Krishnan T, McLachlan GJ. The EM algorithm. In Gentle J, Hardle W, Mori Y, editors. Handbook of computational statistics, vol. 1, Chapter II.5. New York: Springer-Verlag; 2004. p. 137—68.

[18] Blake CL, Merz CJ. UCI Repository of machine learning databases. Irvine, CA: University of California, Department of Information and Computer Science; 1998. Available: http://www.ics.uci.edu/∼mlearn/MLRepository.html (accessed: 1 June, 2007).

[19] Schwarz G. Estimating the dimension of a model. Ann Stat 1978;6:461—4.

[20] Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data, 2nd ed, New Jersey: Wiley; 2002.

[21] Ambroise C, McLachlan GJ. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc Natl Acad Sci USA 2002;99:6562—6.

[22] Lenk PJ, DeSarbo WS. Bayesian inference for finite mixture of generalized linear models with random effects. Psychometrika 2000;65:93—119.

[23] Vermunt JK, Magidson J. Latent class models for classification. Comput Stat Data Anal 2003;41:531—7.

[24] Fruhwirth-Schnatter S, Tuchler R, Otter T. Bayesian analysis of the heterogeneity model. J Bus Econ Stat 2004;22:2—15.

[25] Xu L, Jordan MI, Hinton GE. An alternative model for mixtures of experts. In: Cowan JD, Tesauro G, Alspector J, editors. Adv Neural Inf Proc Systems 7. Cambridge: MIT Press; 1995. p. 633—40.

[26] McLachlan GJ, Krishnan T. The EM algorithm and extensions. New York: Wiley; 1997.