In: Machine Learning Research Progress
Editor: Hanna Peters and Mia Vogel, pp. 1-14
ISBN 989-1-60456-646-8

Chapter 1

EXPERT NETWORKS WITH MIXED CONTINUOUS AND CATEGORICAL FEATURE VARIABLES: A LOCATION MODELING APPROACH

Shu-Kay Ng[1] ***and Geoffrey J McLachlan***[2]
[1]School of Medicine, Logan campus, Griffith University
Meadowbrook, QLD 4131, Australia
[2]Department of Mathematics and Institute for Molecular Bioscience
University of Queensland, Brisbane, QLD 4072, Australia

Abstract

In the context of medically relevant artificial intelligence, many real-world problems involve both continuous and categorical feature variables. When the data are mixed mode, the assumption of multivariate Gaussian distributions for the gating network of normalized Gaussian (NG) expert networks, such as NG mixture of experts (NGME), becomes invalid. An independence model has been studied to handle mixed feature data within the framework of NG expert networks. This method is based on the NAIVE assumption that the categorical variables are independent of each other and of the continuous variables. While this method performs surprisingly well in practice as a way of handling problems with mixed feature variables, the independence assumption is likely to be unrealistic for many practical problems.

In this chapter, we investigate a dependence model which allows for some dependence between the categorical and continuous variables by adopting a location modeling approach. We show how the expectation-maximization (EM) algorithm can still be adopted to train the location NG expert networks via the maximum likelihood (ML) approach. With the location model, the categorical variables are uniquely transformed to a single multinomial random variable with cells of distinct patterns (locations). Any associations between the original categorical variables are then converted into relationships among the resulting multinomial cell probabilities. In practice, the dependence model approach becomes intractable when the multinomial distribution replacing the categorical variables has many cells and/or there are many continuous feature variables. An efficient procedure is developed to determine the correlation structure between the categorical and continuous variables in order to minimize the number of parameters in the dependence model. The method is applied to classify cancer patients on the basis of continuous gene-expression-profile vector of tumour samples and categorical variables of patient's clinical characteristics. The proposed

methodologies would have wide application in various scientific fields such as economy, biomedical and health sciences, and many others, where data with mixed feature variables are collected. Further extensions of the methodologies to other NG networks and/or to other members of the exponential family of densities for the local output density are discussed.

1 INTRODUCTION

Among the various kinds of expert networks, NG expert networks are of much interest due to their wide applicability [1-3] and the advantage of fast learning via the EM algorithm of Dempster et al. [4] without the requirement of a carefully selected learning rate in the inner loop of the EM algorithm [1, 5-6]. Normalized Gaussian expert networks softly partition the input space into, say M, regions by NG functions (the gating network)

$$\mathcal{N}_h(\boldsymbol{x}) = \frac{\pi_h f_h(\boldsymbol{x})}{\sum_{l=1}^{M} \pi_l f_l(\boldsymbol{x})} \qquad (h = 1, \ldots, M), \tag{1}$$

where $\pi_h > 0$, $\sum_{h=1}^{M} \pi_h = 1$, and $f_h(\boldsymbol{x}) = \phi_h(\boldsymbol{x}; \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)$ denotes the multivariate Gaussian function for input vector $\boldsymbol{x}$, with mean $\boldsymbol{\mu}_h$ and covariance matrix $\boldsymbol{\Sigma}_h$. Each local unit (expert) approximates the output within the partition and the final output of the NG network is given by the summation of these local outputs weighted by the normalized Gaussian functions $\mathcal{N}_h(\boldsymbol{x})$ $(h = 1, \ldots, M)$. The architecture is thus based on the divide-and-conquer principle where a complex task is broken up into simpler and smaller subtasks and their solutions can be combined to yield a solution to the complex problem [6]; see, for example, [7].

In the context of medically relevant artificial intelligence, many real-world problems involve both continuous and categorical feature variables. These include learning problems in wide areas of biomedical and health sciences. An example in the context of cancer research is to classify patients based on continuous measurements such as size of primary tumour, as well as categorical variables such as cardiovascular disease history and bone metastases. Precisely, the input vector $\boldsymbol{x}_j$ on the j-th entity consists of q categorical variables in the vector $\boldsymbol{x}_{1j}$ in addition to p continuous variables represented by the vector $\boldsymbol{x}_{2j}$ for $j = 1, \ldots, n$, where n is the total number of observations. When the data are mixed mode, the assumption of multivariate Gaussian distributions for the gating network (1) becomes invalid. An attempt has been studied by Everitt [8] in which the categorical variables $\boldsymbol{x}_{1j}$ are assumed to have arisen through thresholding of unobservable continuous variables. The thresholds that define the categories are treated as extra parameters. The unobserved and observed continuous variables are assumed to be jointly multivariate Gaussian. In practice, the method is limited to one or two categorical variables [9] because the log likelihood contains q-dimensional integrals and is therefore numerically intractable for large q [10]. Recently, an independence model has been studied by Ng and McLachlan [3] to handle mixed feature data within the framework of NG expert networks. This method is based on the NAIVE assumption that the categorical variables are independent of each other and of the continuous variables. Under this independence assumption, $f_h(\boldsymbol{x})$ in (1) can be written

as

$$f_h(\boldsymbol{x}) = f_h(\boldsymbol{x}_1) f_h(\boldsymbol{x}_2|\boldsymbol{x}_1) = \prod_{i=1}^{q} f_{hi}(x_{1i}) \phi_h(\boldsymbol{x}_2; \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h), \tag{2}$$

where $f_{hi}(x_{1i})$ denotes the h-th conditional density of the i-th categorical variable x_{1i} in $\boldsymbol{x}_1$ for $i = 1, \ldots, q$. In (2), the symbol f_h is being used generically to denote a density where, for categorical random variables, the density is really a probability function. While this independence assumption approach performs surprisingly well in practice as a way of handling problems with mixed feature variables [3, 11], the independence assumption is likely to be unrealistic for many practical problems.

In this chapter, we investigate a model which allows for some dependence between the categorical variables (vector $\boldsymbol{x}_{1j}$) and the continuous variables (vector $\boldsymbol{x}_{2j}$) by adopting a location modeling approach, used in discriminant analysis and graphical modeling of mixed variables [9, 12]. With the location model, the categorical variables are uniquely transformed to a single multinomial random variable with cells of distinct patterns (locations). The conditional distribution of the continuous variables, $f_h(\boldsymbol{x}_2|\boldsymbol{x}_1)$ in (2), is taken to be multivariate Gaussian with a mean $\boldsymbol{\mu}_h$ that is allowed to be different for some or all of the distinct patterns of the categorical variables $\boldsymbol{x}_1$. Our aims are to create alternative methodologies in tackling problems with mixed feature data and, in particular, a wider applicability of NG expert networks by incorporating the location model within networks modeling. The rest of the chapter is organized as follows: Section 2 introduces the extension of the NG gating network using the NAIVE independence and location models for problems with mixed feature data. In Section 3, we describe the generalized NGME network and show how the EM algorithm can still be adopted to train the generalized NGME networks via the ML approach. In practice, the location model approach becomes intractable when the multinomial distribution replacing the categorical variables has many cells and/or there are many continuous feature variables. In Section 4, an efficient procedure is developed to determine the correlation structure between the categorical and continuous variables in order to minimize the number of parameters in the dependence model. The proposed methodologies are illustrated in Section 5 using a real example of classifying cancer patients on the basis of continuous gene-expression-profile vector of tumor samples and categorical variables of patient's clinical characteristics. Section 6 ends the paper with some discussion and conclusions.

2 INDEPENDENCE AND LOCATION MODELS

For problems with both categorical and continuous feature variables, we let $\boldsymbol{x}_j = (\boldsymbol{x}_{1j}^T, \boldsymbol{x}_{2j}^T)^T$ denote the feature vector on the j-th entity, where $\boldsymbol{x}_{2j}$ contains the continuous features and $\boldsymbol{x}_{1j} = (x_{11j}, \ldots, x_{1qj})^T$ contains the q categorical variables. Here x_{1ij} is the value of the i-th categorical variable on the j-th entity ($i = 1, \ldots, q$; $j = 1, \ldots, n$), taking on n_i distinct values, and the superscript T denotes vector transpose. The h-th conditional density of the i-th categorical variable is given by a multinomial distribution consisting of one draw on n_i values with probabilities $\lambda_{hi1}, \ldots, \lambda_{hin_i}$, where $\lambda_{hin_i} = 1 - \sum_{l=1}^{n_i - 1} \lambda_{hil}$.

That is, we have

$$f_{hi}(x_{1ij}) = \prod_{v=1}^{n_i} \lambda_{hiv}^{\delta(x_{1ij},v)}, \tag{3}$$

where $\delta(x_{1ij}, v) = 1$ if $x_{1ij} = v$ and is zero otherwise $(v = 1, \ldots, n_i)$. Under the NAIVE independence assumption for the q categorical variables, it follows from (2) and (3) that

$$f_h(\boldsymbol{x}_j) = \prod_{i=1}^{q} \prod_{v=1}^{n_i} \lambda_{hiv}^{\delta(x_{1ij},v)} \phi_h(\boldsymbol{x}_{2j}; \boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h). \tag{4}$$

Although the independence assumption is likely to be unrealistic for many problems, it often performs surprisingly well in practice as a way of handling problems with mixed feature data [3, 11, 13]. One important reason is that the NAIVE method usually requires fewer parameters to be estimated than more complicated alternative methods that try to model interactions between the categorical and continuous variables. The independence model will therefore tend to have a lower variance for the estimates and compensate for any increase in the estimation bias [11]. For problems where the minimization of estimation bias is not the main objective, such as in the cluster analysis, the independence model may provide a better classification result when small sample sizes are involved [11].

The NAIVE model can be modified to allow for some dependence between $\boldsymbol{x}_1$ and the vector $\boldsymbol{x}_2$ of continuous variables by adopting the location model as, for example, in [9, 14]. With the location model, the q categorical variables are uniquely transformed to a single multinomial random variable $\boldsymbol{U}$ with S cells, where $S = \prod_{i=1}^{q} n_i$ is the number of distinct patterns (locations) of the q categorical variables. We denote $(\boldsymbol{u}_j)_s$ the label for the s-th location of the j-th entity $(s = 1, \ldots, S;\ j = 1, \ldots, n)$ and $(\boldsymbol{u}_j)_s = 1$ if the q categorical variables in $\boldsymbol{x}_{1j}$ correspond to the s-th pattern. Any associations between the original categorical variables are then converted into relationships among the resulting multinomial cell probabilities. The location model assumes further that conditional on $(\boldsymbol{u}_j)_s = 1$, the conditional distribution of the p continuous variables $\boldsymbol{x}_{2j}$ is Gaussian with mean $\boldsymbol{\mu}_{hs}$ and covariance matrix $\boldsymbol{\Sigma}_h$; that is, $\phi_h(\boldsymbol{x}_{2j}; \boldsymbol{\mu}_{hs}, \boldsymbol{\Sigma}_h)$, where the covariance matrix is the same for all S cells. Let p_{hs} be the probability that $(\boldsymbol{U}_j)_s = 1$ for the h-th experts $(h = 1, \ldots, M;\ s = 1, \ldots, S)$. The density function $f_h(\boldsymbol{x}_j)$ in (2) is replaced by

$$f_h(\boldsymbol{x}_j) = \prod_{s=1}^{S} \left[p_{hs} \phi_h(\boldsymbol{x}_{2j}; \boldsymbol{\mu}_{hs}, \boldsymbol{\Sigma}_h)\right]^{\delta(j,s)}, \tag{5}$$

where $\delta(j, s) = 1$ if $\boldsymbol{x}_{1j}$ corresponds to the s-th pattern; that is, $(\boldsymbol{u}_j)_s = 1$, and is zero otherwise. In contrast to the thresholding approach considered by Everitt [8], the location model does not impose any orders of the categories in each categorical variable and any structure on the conditional means [10].

3 GENERALIZED NGME NETWORKS

As shown in Figure 1, the (generalized) NGME architecture is comprised of M expert networks. These expert networks approximate the distribution of the output $\boldsymbol{y}_j$ within each region of the input space. The expert network maps its input $\boldsymbol{x}_j$ to a local output, the

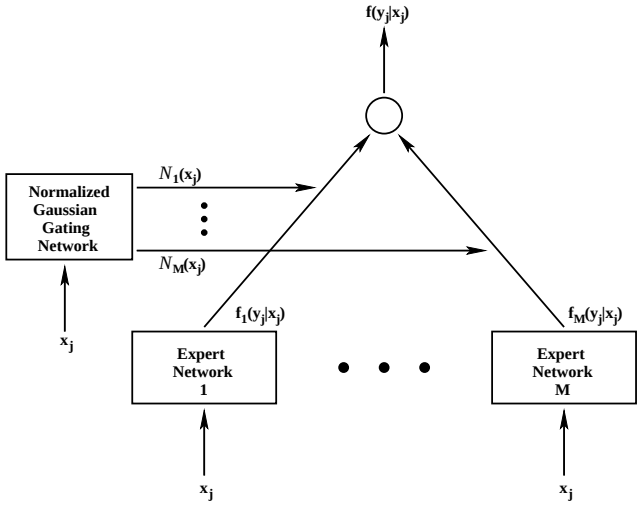

Figure 1: The (generalized) normalized Gaussian mixture of experts

density $f_h(\boldsymbol{y}_j|\boldsymbol{x}_j;\boldsymbol{\theta}_h)$, 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 modeled by NG functions $\mathcal{N}_h(\boldsymbol{x})$ in (1) provides a set of scalar coefficients that weight the contributions of the various experts. These NG functions are now denoted by $\mathcal{N}_h(\boldsymbol{x}_j;\boldsymbol{\pi},\boldsymbol{\alpha})$, where $\boldsymbol{\pi}$ is the symbol for the collection of $\pi_1,\ldots,\pi_{M-1}$ and $\boldsymbol{\alpha}$ denotes a vector of unknown parameters in the density functions $f_h(\boldsymbol{x})$ $(h = 1,\ldots,M)$ in (1). The final output of the NGME neural network is a weighted sum of all the local output vectors produced by expert networks:

$$f(\boldsymbol{y}_j|\boldsymbol{x}_j;\ \boldsymbol{\Psi}) = \sum_{h=1}^{M} \mathcal{N}_h(\boldsymbol{x}_j;\boldsymbol{\pi},\boldsymbol{\alpha}) f_h(\boldsymbol{y}_j|\boldsymbol{x}_j;\boldsymbol{\theta}_h), \tag{6}$$

where $\boldsymbol{\Psi}$ is the vector of all the unknown parameters. The local output densities $f_h(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}_h)$ $(h = 1,\ldots,M)$ can be generally assumed to belong to the exponential family of densities, such as the Gaussian and the Bernoulli distributions, respectively, for regression and binary classification problems [6]. The unknown parameter vector $\boldsymbol{\Psi}$ can be estimated by the ML approach via the EM algorithm [1, 15]. In contrast to the ME networks [16], the learning of NGME networks do not require both the selection of a learning rate and the iterative inner loop in the EM algorithm [1, 5, 6].

To apply the EM algorithm to the generalized NGME networks, we introduce the indicator variables z_{hj}, where z_{hj} is one or zero according to whether $\boldsymbol{y}_j$ belongs or does not belong to the hth expert [5, 17]. We let the missing data $\boldsymbol{z}$ be the vector containing all these indicator variables. Based on an asymmetrical representation for the joint density $f(\boldsymbol{y},\boldsymbol{x}) = f(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\Psi})f(\boldsymbol{x};\boldsymbol{\pi},\boldsymbol{\alpha})$ described in [1, 18], where $f(\boldsymbol{x};\boldsymbol{\pi},\boldsymbol{\alpha}) = \sum_{h=1}^{M}\pi_h f_h(\boldsymbol{x};\boldsymbol{\alpha}_h)$, the complete-data log likelihood for $\boldsymbol{\Psi}$ is given by

$$\log L_c(\boldsymbol{\Psi}) = \sum_{j=1}^{n}\sum_{h=1}^{M} z_{hj}\{\log\pi_h + \log f_h(\boldsymbol{x}_j;\boldsymbol{\alpha}_h) + \log f_h(\boldsymbol{y}_j|\boldsymbol{x}_j;\boldsymbol{\theta}_h)\}. \tag{7}$$

It follows on application of the EM algorithm in training generalized NGME networks that on the $(k+1)$th iteration, the E-step calculates the Q-function as

$$\begin{aligned} Q(\boldsymbol{\Psi};\, \boldsymbol{\Psi}^{(k)}) &= E_{\boldsymbol{\Psi}^{(k)}}\{\log L_c(\boldsymbol{\Psi})|\boldsymbol{y},\, \boldsymbol{x}\} \\ &= \sum_{j=1}^{n}\sum_{h=1}^{M} E_{\boldsymbol{\Psi}^{(k)}}(Z_{hj}|\boldsymbol{y},\boldsymbol{x})\{\log \pi_h + \log f_h(\boldsymbol{x}_j;\boldsymbol{\alpha}_h) \\ &\qquad + \log f_h(\boldsymbol{y}_j|\boldsymbol{x}_j;\boldsymbol{\theta}_h)\}. \end{aligned} \tag{8}$$

It can be seen that the complete-data log likelihood (8) is linear in $\boldsymbol{z}$. Thus, the E-step just replaces z_{hj} in (7) by its current conditional expectation $\tau_{hj}^{(k)}$ given $\boldsymbol{y}_j$, $\boldsymbol{x}_j$, and the current estimate $\boldsymbol{\Psi}^{(k)}$ for $\boldsymbol{\Psi}$, where

$$\begin{aligned} \tau_{hj}^{(k)} &= \mathrm{pr}_{\boldsymbol{\Psi}^{(k)}}\{Z_{hj}=1|\boldsymbol{y}_j,\boldsymbol{x}_j\} \\ &= \frac{\pi_h^{(k)} f_h(\boldsymbol{x}_j;\boldsymbol{\alpha}_h^{(k)}) f_h(\boldsymbol{y}_j|\boldsymbol{x}_j;\boldsymbol{\theta}_h^{(k)})}{\sum_{l=1}^{M}\pi_l^{(k)} f_l(\boldsymbol{x}_j;\boldsymbol{\alpha}_l^{(k)}) f_l(\boldsymbol{y}_j|\boldsymbol{x}_j;\boldsymbol{\theta}_l^{(k)})} \end{aligned} \tag{9}$$

for $h=1,\ldots,M$. From (8), it can be seen that the Q-function can be decomposed into three terms with respect to π_h, $\boldsymbol{\alpha}_h$, and $\boldsymbol{\theta}_h$ $(h=1,\ldots,M)$, respectively. That is,

$$Q_\pi = \sum_{j=1}^{n}\sum_{h=1}^{M} \tau_{hj}^{(k)} \log \pi_h, \tag{10}$$

$$Q_\alpha = \sum_{j=1}^{n}\sum_{h=1}^{M} \tau_{hj}^{(k)} \log f_h(\boldsymbol{x}_j;\boldsymbol{\alpha}_h), \tag{11}$$

and

$$Q_\theta = \sum_{j=1}^{n}\sum_{h=1}^{M} \tau_{hj}^{(k)} \log f_h(\boldsymbol{y}_j|\boldsymbol{x}_j;\boldsymbol{\theta}_h). \tag{12}$$

The decomposition of the Q-function implies that the updated estimates of $\boldsymbol{\pi}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\theta}$ can be obtained in the M-step by maximizing the corresponding decomposed Q-functions separately.

3.1 Learning algorithm for the independence model

With the independence model described in Section 2, the vector of unknown parameters $\boldsymbol{\alpha}_h$ for $f_h(\boldsymbol{x})$ in (4) consists of λ_{hiv} $(i=1,\ldots,q;\, v=1,\ldots,n_i-1)$, and the elements of $\boldsymbol{\mu}_h$ and $\boldsymbol{\Sigma}_h$ $(h=1,\ldots,M)$. The E-step involves the calculation of $\tau_{hj}^{(k)}$ in (9) with $f_h(\boldsymbol{x}_j;\boldsymbol{\alpha}_h^{(k)})$ replaced by (4) based on the current estimate $\boldsymbol{\alpha}_h^{(k)}$. In the M-step, $\pi_h^{(k+1)}$ is obtained by maximizing Q_π in (10) as

$$\pi_h^{(k+1)} = \sum_{j=1}^{n} \tau_{hj}^{(k)}/n. \tag{13}$$

Similarly, $\boldsymbol{\alpha}_h^{(k+1)}$ is obtained by maximizing Q_α in (11). Based on (4), we have

$$\lambda_{hiv}^{(k+1)} = \frac{\sum_{j=1}^n \tau_{hj}^{(k)} \delta(x_{1ij}, v)}{\sum_{j=1}^n \tau_{hj}^{(k)}}. \tag{14}$$

It is noted that (14) can be modified slightly to limit the effect of zero estimates of λ_{hiv} for rare values v as

$$\lambda_{hiv}^{(k+1)} = \frac{\sum_{j=1}^n \tau_{hj}^{(k)} \delta(x_{1ij}, v) + 1/n_i}{\sum_{j=1}^n \tau_{hj}^{(k)} + 1}, \tag{15}$$

by taking into account the number of possible categories n_i in the i-th categorical variable x_{1i} $(i = 1, \ldots, q)$; see [11, 13]. From (4) and (11), the updates of the means $\boldsymbol{\mu}_h$ and covariance matrices $\boldsymbol{\Sigma}_h$ for $h = 1, \ldots, M$ are given by

$$\boldsymbol{\mu}_h^{(k+1)} = \frac{\sum_{j=1}^n \tau_{hj}^{(k)} \boldsymbol{x}_{2j}}{\sum_{j=1}^n \tau_{hj}^{(k)}} \tag{16}$$

and

$$\boldsymbol{\Sigma}_h^{(k+1)} = \frac{\sum_{j=1}^n \tau_{hj}^{(k)} (\boldsymbol{x}_{2j} - \boldsymbol{\mu}_h^{(k+1)})(\boldsymbol{x}_{2j} - \boldsymbol{\mu}_h^{(k+1)})^T}{\sum_{j=1}^n \tau_{hj}^{(k)}}. \tag{17}$$

Depending on the local output densities specified for $f_h(\boldsymbol{y}|\boldsymbol{x}; \boldsymbol{\theta}_h)$, the updated estimate of $\boldsymbol{\theta}_h^{(k+1)}$ is obtained by solving

$$\sum_{j=1}^n \tau_{hj}^{(k)} \partial \log f_h(\boldsymbol{y}_j|\boldsymbol{x}_j; \boldsymbol{\theta}_h)/\partial \boldsymbol{\theta}_h = \mathbf{0} \tag{18}$$

for each h $(h = 1, \ldots, M)$. For example, $f_h(y_j|\boldsymbol{x}_j; \boldsymbol{\theta}_h)$ are assumed to be Gaussian as

$$f_h(y_j|\boldsymbol{x}_j; \boldsymbol{\theta}_h) = \frac{1}{\sqrt{(2\pi\sigma_h^2)}} \exp\{-\tfrac{1}{2}(y_j - \boldsymbol{w}_h^T \boldsymbol{x}_j)^2/\sigma_h^2\}, \tag{19}$$

where $\boldsymbol{w}_h$ and σ_h^2 are, respectively, the weight vector and the variance (dispersion parameter) of the h-th expert network. For notational convenience, we still present the mixed-mode input vector as $\boldsymbol{x}_j$ in (19). Indeed, the categorical variables are replaced by $n_i - 1$ dummy variables and contribute to the local output via the linear predictor $\eta_{hj} = \boldsymbol{w}_h^T \boldsymbol{x}_j$; see [5]. From (18), the updates of $\boldsymbol{\theta}_h^T = (\boldsymbol{w}_h^T, \sigma_h^2)$ for $h = 1, \ldots, M$ are given by

$$\boldsymbol{w}_h^{(k+1)} = \left[\sum_{j=1}^n \tau_{hj}^{(k)} \boldsymbol{x}_j \boldsymbol{x}_j^T\right]^{-1} \sum_{j=1}^n \tau_{hj}^{(k)} y_j \boldsymbol{x}_j \tag{20}$$

and

$$\sigma_h^{(k+1)} = \frac{\sum_{j=1}^n \tau_{hj}^{(k)} (y_j - \boldsymbol{w}_h^{(k+1)^T} \boldsymbol{x}_j)^2}{\sum_{j=1}^n \tau_{hj}^{(k)}}. \tag{21}$$

For binary classification problems, the local densities $f_h(y_j|\boldsymbol{x}_j;\boldsymbol{\theta}_h)$ are assumed to be Bernoulli distribution of possible binary outcomes of "failure" and "success" [6]. That is,

$$f_h(y_j|\boldsymbol{x}_j;\boldsymbol{\theta}_h) = \left(\frac{\exp(\boldsymbol{w}_h^T\boldsymbol{x}_j)}{1+\exp(\boldsymbol{w}_h^T\boldsymbol{x}_j)}\right)^{y_j}\left(\frac{1}{1+\exp(\boldsymbol{w}_h^T\boldsymbol{x}_j)}\right)^{1-y_j}, \tag{22}$$

where $\boldsymbol{\theta}_h = \boldsymbol{w}_h$. In this case, equation (18) becomes

$$\sum_{j=1}^{n}\tau_{hj}^{(k)}\left(y_j - \frac{\exp(\boldsymbol{w}_h^T\boldsymbol{x}_j)}{1+\exp(\boldsymbol{w}_h^T\boldsymbol{x}_j)}\right)\boldsymbol{x}_j = \mathbf{0} \tag{23}$$

for $h = 1,\ldots,M$, which are M sets of nonlinear equations each with unknown parameter vector $\boldsymbol{w}_h$ [3, 5].

3.2 Learning algorithm for the location model

With the location model described in Section 2, the vector of unknown parameters $\boldsymbol{\alpha}_h$ for $f_h(\boldsymbol{x})$ in (5) consists of p_{hs}, and the elements of $\boldsymbol{\mu}_{hs}$ and $\boldsymbol{\Sigma}_h$ $(h = 1,\ldots,M;\ s = 1,\ldots,S)$. The E-step involves the calculation of $\tau_{hj}^{(k)}$ in (9) with $f_h(\boldsymbol{x}_j;\boldsymbol{\alpha}_h^{(k)})$ replaced by (5) based on the current estimate $\boldsymbol{\alpha}_h^{(k)}$. In the M-step, both $\pi_h^{(k+1)}$ and $\boldsymbol{\theta}_h^{(k+1)}$ are obtained according to (13) and (18) in Section 3.1, respectively. For the updated estimate of $\boldsymbol{\alpha}_h^{(k+1)}$, we have, from (5) and (11),

$$p_{hs}^{(k+1)} = \frac{\sum_{j=1}^{n}\tau_{hj}^{(k)}\delta(j,s)}{\sum_{j=1}^{n}\tau_{hj}^{(k)}}, \tag{24}$$

$$\boldsymbol{\mu}_{hs}^{(k+1)} = \frac{\sum_{j=1}^{n}\tau_{hj}^{(k)}\delta(j,s)\boldsymbol{x}_{2j}}{\sum_{j=1}^{n}\tau_{hj}^{(k)}\delta(j,s)}, \tag{25}$$

and

$$\boldsymbol{\Sigma}_h^{(k+1)} = \frac{\sum_{j=1}^{n}\tau_{hj}^{(k)}\sum_{s=1}^{S}\delta(j,s)(\boldsymbol{x}_{2j}-\boldsymbol{\mu}_{hs}^{(k+1)})(\boldsymbol{x}_{2j}-\boldsymbol{\mu}_{hs}^{(k+1)})^T}{\sum_{j=1}^{n}\tau_{hj}^{(k)}}. \tag{26}$$

for $h = 1,\ldots,M$ and $s = 1,\ldots,S$.

4 CORRELATION STRUCTURE BETWEEN CATEGORICAL AND CONTINUOUS VARIABLES

In practice, the number of parameters with the location model approach can be large if the mulitnomial distribution replacing the categorical variables has many cells and there are several continuous feature variables. Thus, to implement the location model on NG networks, the intent is to examine preliminary fits (such as correlations, scatter plots, and two-way tables). If strong associations between two categorical variables are detected from preliminary fits, then the two variables may be combined into a single multinomial variable

Table 1: Original coding of the binary variables (breast cancer data)

Variable	Value=0	Value=1
tumor grade (GR)	grades 1 or 2	grade 3
oestrogen receptor status (ER)	$\leq$10	$>$10
progesteron receptor status (PR)	$\leq$10	$>$10
tumor size (SZ)	$\leq$20mm	$>$20mm
patient age (AG)	$\leq$40	$>$40
angioinvasion (AN)	no	yes

with a cell for each category of the two-way table (or fewer, if some categories are pooled); see, for example, [14]. The correlation between continuous variables can also be examined to detect the existence of any relationship among the continuous variables. If the dependency among covariates is weak, then a diagonal covariance matrix for $\mathbf{\Sigma}_h$ may be adopted to reduce the number of unknown parameters for the location model [14].

The simplest location model is the NAIVE-location model, where $f_h(\boldsymbol{x}_2|\boldsymbol{x}_1)$ in (2) is taken to be multivariate Gaussian with a mean that depends only on one of the q categorical variables. This categorical variable can be determined based on the significance of testing the difference in the continuous variables between different categories of the categorical variable. A high level of significance justifies the interaction between the categorical variable and the vector x_2 of continuous variables for the location model. Alternatively, additional local associations between the variables, categorical or continuous or both, can be included to expand the NAIVE-location model on the basis of the likelihood ratio test that measures the change in the log likelihood.

5 ILLUSTRATION: CLASSIFICATION OF BREAST CANCER PATIENTS

We illustrate the proposed methodologies using a real example of classifying breast cancer patients on the basis of the gene expression-profile vector of tumor samples and categorical variables of patient's clinical characterisitcs. The original data set [19] consists of 5000 gene expression profiles and 6 binary variables of clinical indicators from 78 sporadic lymph-node-negative breast cancer patients. With these patients, 44 remained metastasis free after a period of more than 5 years (good prognosis) and 34 patients had developed distant metastases within 5 years (poor prognosis). Based only on the 5000 gene expression profiles, van't Veer et al. [19] identified 70 genes that are associated with disease outcome (distant metastases within 5 years) of cancer patients. Alternatively, the clustering of genes on the basis of gene expression profiles can be employed to form a smaller number of subgroups of genes [20, 21]. Each subgroup of genes is then represented by a single vector (a "metagene") for the subsequent clustering of the tissue samples [22] and the identification of biological "markers" for disease outcome [23].

In this study, we work on the data set with 6 binary variables of clinical indicators and

Table 2: Model selection (breast cancer data)

Number of experts	Worth indices
2	(0.78, 0.22)*
3	(0.73, 0.18, 0.09)

* The number of experts selected.

5 continuous variables representing the top 5 metagenes ranked in terms of the likelihood ratio statistic described in [22]; see also [3]. Table 1 displays the original coding of the 6 binary clinical indicators given in the Supplementary Information of [19]. We first apply the NGME network of [1] on the continuous variables to classify the patients into good and poor prognosis subgroups; see equation (22). This preliminary analysis provides the initial estimates and the determination of the number of experts M for the generalized NGME network. In addition, the improvement of the generalized NGME network by using additional binary clinical indicators can be assessed. Such evaluation is based on the misclassification error rate using the "leave-one-out" method for cross-validation. The number of experts M is determined based on a frequentist analog of the "worth index" on model selection. The worth index for the h-th expert, based on the indicator variables z_{hj} over the data, is given in [24] as

$$I_h = \sum_{j=1}^{n} z_{hj}/n \qquad (h = 1, \ldots, M).$$

Here, we consider a frequentist analog where z_{hj} is replaced by its estimated conditional expectation $\hat{\tau}_{hj}$. The number of experts is chosen to be the minimum value of M with the largest worth indices for which the sum of their worth indices exceeds 0.8; see [24, 25]. Table 2 displays the results for the model selection. A NGME network with $M = 2$ experts is selected. The leave-one-out error rate is provided in Table 3. We then apply the generalized NGME networks to classify the patients into good and poor prognosis subgroups, using the independence and location models, on the mixed feature data. The results for these two methods are displayed in Table 3. With the independence model, the results using (14) and (15) for the update of $\lambda_{hiv}^{(k+1)}$ are the same. With the location model, we consider a NAIVE-location model described in Section 4 where the conditional means are specified to each level (location) of the combined variable of the oestrogen receptor (ER) and progesteron receptor (PR) status, as a very strong association between these two categorical variables is detected; see Table 4. It is also observed that there is highly significant difference (p-value < 0.0005) in all continuous variables between different categories of this combined variable $\{$ER, PR$\}$. This indicates a significant interaction between $\{$ER, PR$\}$ and the vector of continuous variables for the NAIVE-location model.

From Table 3, it can be seen that the generalized NGME networks significantly reduce the error rate by using additional binary clinical indicators. There is, however, little difference between the independence and location models for the binary classification of this breast cancer data. The allocation of two patients out of 78 is corrected using the location model. Although there are significant correlations between categorical variables and inter-

Table 3: Leave-one-out error rates (breast cancer data)

Method	Error rate
NGME network on continuous variables	29.5%
Independence model on mixed variables	19.2%
Location model on mixed variables	16.7%

Table 4: Association between categorical variables (breast cancer data): Chi-square statistic with the p-value in brackets

Variables	ER	PR	SZ	AG	AN
GR	9.6 (0.002)	11.8 (0.001)	9.7 (0.002)	3.5 (0.060)	2.3 (0.129)
ER		39.7 (0.000)	6.7 (0.009)	0.0 (0.965)	0.5 (0.500)
PR			7.0 (0.002)	0.0 (0.937)	0.1 (0.765)
SZ				0.7 (0.418)	1.2 (0.282)
AG					0.4 (0.541)

actions between categorical and continuous variables, the performance of the independence model is comparable to that of the location model for the binary classification of the breast cancer data. As presented in Section 2, the independence model tends to have a lower variance for the estimates and compensates for any increase in the estimation bias, especially when the sample size is small.

6 CONCLUSION

Many kinds of data collected in wide areas of machine learning applications involve both categorical and continuous feature variables. With the mixed feature data, the assumption of multivariate Gaussian becomes invalid with NG expert networks. In this chapter, we have extended the NGME network to incorporate the independence and location models for tackling problems with mixed feature data. The independence model assumes that the categorical variables are independent of each other and of the continuous variables. The location model extends the independence model by allowing the possibility of within-expert associations between categorical and continuous variables. These methodologies provide alternative methods in networks modeling within the framework of NG networks with mixed feature data.

Normalized Gaussian ME networks with NG gating networks have the advantage of efficient learning via the EM algorithm [1, 18], which has a number of desirable properties including its numerically stability and reliable convergence [17]. This is in contrast to the ME networks [6, 16], where the nonlinearity of the *softmax* gating network [26] implies that an iterative reweighted least squares (IRLS) algorithm with a carefully selected learning

rate is required in the inner loop of the EM algorithm [5, 6]. We show in Section 3 that the desirable property of fast learning via the EM algorithm is preserved within the generalized NGME networks incorporating the independence and location models for problems with mixed feature data. In practice, the location modeling approach becomes intractable when the multinomial distribution replacing the categorical variables has many cells and/or there are many continuous feature variables. We have developed an efficient procedure in Section 4 to determine the correlation structure between the categorical and continuous variables in order to minimize the number of parameters in the location model.

In Section 5, the proposed methods are illustrated using a real example in the context of bioinformatics. From Table 3, it can be seen that significant improvement is achieved by using additional categorical variables via the independence or location models. The error rates in Table 3 have been considered here in a relative sense. However, caution should be exercised in interpreting these rates in an absolute sense. This is because the metagenes in the data set are determined using the expression profiles from the 78 cancer patients. Thus, the misclassification error rate is calculated without allowance for the selection bias, which is present because each "leave-one-out" test sample was also used in the gene-selection process [3, 27]. The error rates given in Table 3 should therefore be interpreted as apparent error rates. An "external" cross-validation can be adopted to correct for the bias in estimating the error of a prediction rule, where gene-selection procedure is performed at each stage of the cross-validation process on the remaining samples; see [20, 27].

The proposed methodologies would have wide application in various scientific fields such as economy, biomedical and health sciences, and many others, where data with mixed feature variables are collected. Although the focus of the chapter is on the NGME network, the methodologies can be readily applied to the NG radial basis function (NGRBF) networks [28, 29], based on the connection between the NGME and NGRBF networks described in [2]. In addition, other members of the exponential family of densities can be adopted for the specification of the local output density $f_h(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\theta}_h)$; see [6]. Jiang and Tanner [30] have obtained conditions for the identifiability of the ME network, which they showed held for some commonly used expert networks such as Poisson, gamma, Gaussian, and Bernoulli experts.

References

[1] Xu, L., Jordan, M.I., & Hinton, G.E. (1995). An alternative model for mixtures of experts. In J.D. Cowan, G. Tesauro, & J. Alspector (Eds.), *Adv. in Neural Inf. Proc. Systems 7* (pp. 633–640). Cambridge, Massachusetts: MIT Press.

[2] Xu, L. (1998). RBF nets, mixture experts, and Bayesian Ying-Yang learning. *Neurocomputing*, 19, 223–257.

[3] Ng, S.K. & McLachlan, G.J. (2005). Normalized Gaussian networks with mixed feature data. *Lecture Notes in Artificial Intelligence*, 3809, 879–882.

[4] Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B*, 39, 1–38.

[5] Ng, S.K. & McLachlan, G.J. (2004). Using the EM algorithm to train neural networks: Misconceptions and a new algorithm for multiclass classification. *IEEE T. Neural Networks*, 15, 738–749.

[6] Jordan, M.I. & Jacobs, R.A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Comput.*, 6, 181–214.

[7] Rao, A.V., Miller, D., Rose, K. & Gersho, A. (1997). Mixture of experts regression modeling by deterministic annealing. *IEEE T. Signal Proces.*, 45, 2811–2820.

[8] Everitt, B.S. (1988). A finite mixture model for the clustering of mixed-mode data *Stat. Probabil. Lett.*, 6, 305–309.

[9] Lawrence, C.J. & Krzanowski, W.J. (1996). Mixture separation for mixed-mode data. *Stat. Comput.*, 6, 85–92.

[10] Willse, A. & Boik, R.J. (1999). Identifiable finite mixtures of location models for clustering mixed-mode data. *Stat. Comput.*, 9, 111–121.

[11] Hand, D.J. & Yu, K.M. (2001). Idiot's Bayes – not so stupid after all? *Int. Stat. Rev.*, 69, 385–398.

[12] Whittaker, J. (1990). Graphical Models in Applied Multivariate Statistics. Chichester: Wiley.

[13] Titterington, D.M., Murray, G.D., Murray, L.S., Spiegelhalter, D.J., Skene, A.M., Habbema, J.D.F., & Gelpke, G.J. (1981). Comparison of discrimination techniques applied to a complex data set of head injured patients. *J. Roy. Stat. Soc. Ser. A*, 144, 145–175.

[14] Hunt, L.A. & Jorgensen, M.A. (1999). Mixture model clustering: a brief introduction to the MULTIMIX program. *Aust. NZ. J. Stat.*, 40, 153–171.

[15] Jordan, M.I. & Xu, L. (1995). Convergence results for the EM approach to mixtures of experts architectures. *Neural Networks*, 8, 1409–1431.

[16] Jacobs, R.A., Jordan, M.I., Nowlan, S.J., & Hinton, G.E. (1991). Adaptive mixtures of local experts. *Neural Comput.*, 3, 79–87.

[17] Ng, S.K., Krishnan, T, & McLachlan, G.J. (2004). The EM algorithm. In J. Gentle, W. Hardle, & Y. Mori (Eds.), *Handbook of Computational Statistics Vol. 1* (pp. 137–168). New York: Springer-Verlag.

[18] Sato, M. & Ishii, S. (2000). On-line EM algorithm for the normalized Gaussian network. *Neural Comput.*, 12, 407–432.

[19] van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A.M., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernards, R., & Friend, S.H. (2002). Gene expression profiling predicts clinical outcomes of breast cancer. *Nature*, 415, 530–536.

[20] McLachlan, G.J., Do, K.A., & Ambroise, C. (2004). Analyzing Microarray Gene Expression Data (Chapters 5-7). Hobokin, New Jersey: Wiley.

[21] Ng, S.K., McLachlan, G.J., Wang, K., Ben-Tovim Jones, L., & Ng, S.-W. (2006). A mixture model with random-effects components for clustering correlated gene-expression profiles. *Bioinformatics*, 22, 1745–1752.

[22] McLachlan, G.J., Bean, R.W., & Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, 18, 413–422.

[23] Ben-Tovim Jones, L., Ng, S.K., Ambroise, C., Monico, K., Khan, N., & McLachlan, G.J. (2005). Use of microarray data via model-based classification in the study and prediction of survival from lung cancer. In J.S. Shoemaker, S.M. Lin (Eds.), *Methods of Microarray Data Analysis IV* (pp. 163–173). New York: Springer.

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

[25] Ng, S.K. & McLachlan, G.J. (2007). Extension of mixture-of-experts networks for binary classification of hierarchical data. *Artificial Intelligence in Medicine*, 41, 57–67.

[26] Bridle, J.S. (1990). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In F.F. Soulié, J. Hérault (Eds.), *Neuro-Computing: Algorithms, Architectures and Applications* (pp. 227–236). Berlin, Germany: Springer.

[27] Ambroise, C. & McLachlan, G.J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. USA*, 99, 6562–6566.

[28] Moody, J. & Darken, C.J. (1989). Fast learning in networks of locally-tuned processing units. *Neural Comput.*, 1, 281–294.

[29] Bugmann, G. (1998). Normalized Gaussian radial basis function networks. *Neurocomputing*, 20, 97–110.

[30] Jiang, W. & Tanner, M.A. (1999). On the identifiability of mixtures-of-experts. *Neural Networks*, 12, 1253–1258.