Quadratically Gated Mixture of Experts for Incomplete Data Classification

Xuejun Liao XJLIAO@EE.DUKE.EDU
Hui Li HL1@EE.DUKE.EDU
Lawrence Carin LCARIN@EE.DUKE.EDU
Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708

Abstract

We introduce *quadratically gated mixture of experts* (QGME), a statistical model for multi-class nonlinear classification. The QGME is formulated in the setting of incomplete data, where the data values are partially observed. We show that the missing values entail joint estimation of the data manifold and the classifier, which allows *adaptive* imputation during classifier learning. The expectation maximization (EM) algorithm is derived for joint likelihood maximization, with adaptive imputation performed analytically in the E-step. The performance of QGME is evaluated on three benchmark data sets and the results show that the QGME yields significant improvements over competing methods.

1. Introduction

Incomplete data arise for a variety of reasons. In psychological studies, a participant's response is incomplete if he refuses to answer certain questions (Schafer & Graham, 2002); in computer vision the observation of an object becomes incomplete if the object is occluded or some sensor fails to operate (Ghahramani & Jordan, 1994). In DNA microarray analysis, incomplete gene expression data may result from image corruptions, dust or scratches on the slides (Troyanskaya et al., 2001). Incomplete data can lead to serious degradation of statistical learning systems if missing values are ignored or handled inappropriately.

A widely used approach to incomplete data analysis is based on imputation, i.e., replacing the missing values with plausible surrogates. Single imputation methods include imputing means or conditional means (Schafer & Graham, 2002). Rubin's multiple imputation (MI) (Rubin, 1987) solves the problem of uncertainty associated with single imputations and yet retains simplicity by allowing complete-data techniques to be applied without modification. The fact that most classification techniques have been designed for complete data makes imputation methods convenient in practice.

Imputation methods ignore the specifics of the algorithms subsequently applied to the imputed data, and thus are unlikely to achieve the full potential offered by those algorithms. The work in (Ibrahim, 1990) circumvents the deficiency of outside-of-algorithm imputations by performing imputation inside the algorithm; since the data are assumed to take finite discrete values, the imputation is intrinsically finite. The work in (Williams et al., 2005) represents an infinite imputation method for continuous data, with missing values handled by analytical integration; the imputation model is estimated separately from the algorithm and the method is limited to linear binary classification.

Nonlinear classification can be accomplished by using nonlinear activations in neural networks or by augmenting linear classifiers with kernel functions. The nonlinear components in these methods, however, make it difficult to handle incomplete data. To make an algorithm amenable to incomplete data, it is helpful to keep its basic components linear.

Given a data set, we can either model it by a single complicated model or by multiple simple models. Linear models are simple and suitable for handling incomplete data, but they are incapable of modeling nonlinear data. Nevertheless, any arbitrary nonlinear models can be approximated by a piecewise linear model, with each linear component modeling a local data manifold where the data are approximately linear.

The hierarchical mixture of experts (HME) in (Jor-

Appearing in *Proceedings of the 24th International Conference on Machine Learning*, Corvallis, OR, 2007.

dan & Jacobs, 1994) can be used as a piecewise linear model. The linear gating nodes, however, entail a hierarchical gating structure, which encumbers handling of incomplete data and makes HME inappropriate for our purpose. The piecewise linear regression (PLR) model in (Li et al., 2006) uses a single-level structure of ellipsoids to form the gating network, which partitions the data space into quadratic subsets. The single-level structure and quadratic nature of the gating nodes greatly facilitates handling of incomplete data.

In this paper we extend the regression model in (Li et al., 2006) to the case of classification, by letting the response variable y take categorical values (class labels) and replacing the linear models with multinomial probit models (Albert & Chib, 1993). The resulting model is termed *Quadratically Gated Mixture of Experts* (QGME). If the latent variable z in probit models are observed, the linear models are recovered and the QGME reduces to the regression model in (Li et al., 2006). We formulate the QGME in the setting of incomplete data, where a feature vector (datum fed to the model) $\mathbf{x}$ is partially observed, i.e., some components in $\mathbf{x}$ are missing. We derive the maximum-likelihood (ML) estimator for EGMEs via expectation maximization (EM). Unlike the model in (Williams et al., 2005), the QGME performs *adaptive* infinite imputation of missing values, implemented in the E-step of the EM algorithm. The adaptivity derives from a joint estimation of the imputation model and the classifier.

2. The QGME Model

Let $\mathbf{x}_i$ be a column vector containing d real-number features, $y_i \in \{0, 1, \cdots, M-1\}$ be the class label associated with $\mathbf{x}_i$, and $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote a multivariate normal distribution of $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Let $\mathbf{I}$ denote an identity matrix. We define

$$p(\zeta_i = k) = \pi_k \tag{1}$$

$$p(\mathbf{x}_i|\zeta_i = k) = \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{2}$$

$$p(y_i|\mathbf{x}_i, \zeta_i = k) = \int_{\mathbf{T}_{y_i}\mathbf{z}_{ik} \le \mathbf{0}} \mathcal{N}(\mathbf{z}_{ik}; \mathbf{W}_k^T \mathbf{x}_i + \mathbf{b}_k, \mathbf{I}) d\mathbf{z}_{ik} \tag{3}$$

from which follows

$$p(\zeta_i = k|\mathbf{x}_i) = \frac{\pi_k \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \tag{4}$$

$$p(y_i|\mathbf{x}_i) = \sum_{k=1}^K p(\zeta_i = k|\mathbf{x}_i) p(y_i|\mathbf{x}_i, \zeta_i = k) \tag{5}$$

We term the model in (5) as *quadratically gated mixture of experts* (QGME), in which $p(y_i|\mathbf{x}_i, \zeta_i = k)$ is the k-th expert, $p(\zeta_i = k|\mathbf{x}_i)$ is a quadratic gating node giving the probability of selecting the k-th expert, and the remaining quantities are explained in the following

- ζ_i – is a latent gating variable indicating which expert is being selected
- π_k – π_k is the prior probability of selecting the k-th expert
- $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ – given $\zeta_i = k$, $\mathbf{x}_i$ is governed by a normal distribution with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$; the posterior of ζ_i given $\mathbf{x}_i$ is given by Bayes rule, which determines the probability of selecting the k-th expert after $\mathbf{x}_i$ is observed
- $\mathbf{W}_k, \mathbf{b}_k$ – are parameters of the k-th expert, which is a multinomial probit, with linear weights $\mathbf{W}_k$ and intercepts $\mathbf{b}_k$
- $\mathbf{z}_{ik}$ – $\mathbf{z}_{ik} = [z_{i,k,1}, \cdots, z_{i,k,M-1}]$ with $z_{i,k,m}$ the latent utility of class m in the k-th expert (probit model); by default, $z_{i,k,0} = 0$, meaning class 0 has a zero utility
- $\mathbf{T}$ – $\mathbf{T}_0$ is a $(M-1) \times (M-1)$ identity matrix; $\mathbf{T}_m$ $(m = 1, 2, \cdots, M-1)$ is equal to $\mathbf{T}_0$ except the elements in the m-th column are all -1; $\mathbf{T}_m \mathbf{z}_{ik} \le \mathbf{0}$ specifies the constraint $\{\mathbf{z}_{ik} : z_{i,k,m} = \max_{0 \le l \le M-1}\{z_{i,k,l}\}\}$, which defines the region in the $\mathbf{z}_{ik}$ space that is allotted to class m

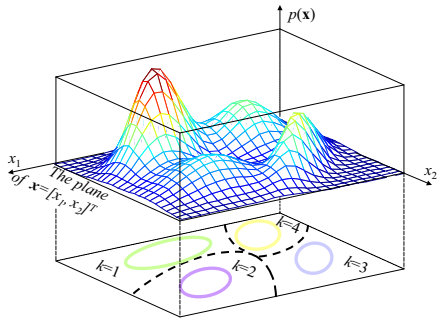

Figure 1. Illustration of the quadratic gating in QGME.

As shown in Figure 1, the space of $\mathbf{x}$ is partitioned probabilistically through a mixture of K normal distributions, whose contours are shown as the ellipsoids. The partition yields a single-level probabilistic gating network, which is used in QGME to select the experts. By assigning any given $\mathbf{x}$ to the normal that yields the maximum probability for $\mathbf{x}$, the space is partitioned into K hard regions and the boundary of each region is defined by quadratic functions. The hard partition yields a deterministic gating network, with each region associated with a unique expert.

The experts in QGME are probit models, each implementing a linear classifier. For a given expert k, we express the conditional class probability $p(y_i|\mathbf{x}_i, \zeta_i = k)$ as the integral of a multivariate normal distribution

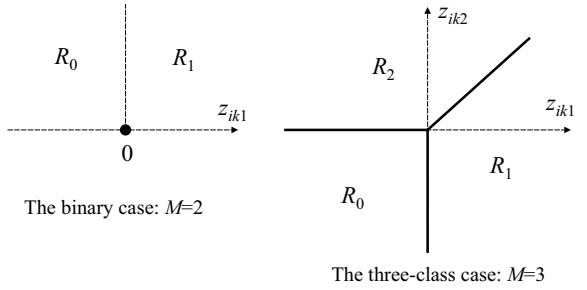

Figure 2. Illustration of the k-th probit expert in QGME, which gives the conditional class probability $p(y_i = m|\mathbf{x}_i, \zeta_i = k) = p(\mathbf{z}_{ik} \in R_m)$, for $m = 0, 1, \cdots, M-1$, where M is the total number of classes, $z_{ik0} \equiv 0$, $\mathbf{z}_{ik}$ is drawn from a multivariate normal distribution with mean $\mathbf{W}_k^T\mathbf{x}_i + \mathbf{b}_k$ and covariance matrix $\mathbf{I}$.

$\mathcal{N}(\mathbf{z}_{ik}; \mathbf{W}_k^T\mathbf{x}_i + \mathbf{b}_k, \mathbf{I})$ over a y-dependent region R_y in the space of $\mathbf{z}_{ik}$, a latent vector introduced for $\mathbf{x}_i$ given $\zeta_i = k$. It must be satisfied that $R_m \cap R_l$ is empty for $m \neq l$ and $\cup_{m=0}^{M-1} R_m$ is the space of $\mathbf{z}_{ik}$. Figure 2 illustrates the R_y used in QGME, which is specified by the linear constraint $\mathbf{T}_y\mathbf{z}_{ik} \leq \mathbf{0}$, where $\mathbf{T}_y$ is defined in the QGME specification above. It is clear from Figure 2 that $\mathbf{z}_{ik} \in R_0$ implies all components of $\mathbf{z}_{ik}$ are less than zero, and $\mathbf{z}_{ik} \in R_y$ $(y = 1, \cdots, M-1)$ implies the y-th component of $\mathbf{z}_{ik}$ is greater than zero as well as other components of $\mathbf{z}_{ik}$. Thus the components of $\mathbf{z}_{ik}$ can be explained as the utilities of classes $1, \cdots, M-1$ and the utility of class 0 is constantly zero (class 0 is the reference class).

It follows from the definition of QGME that the probability of $\mathbf{x}_i$ is a mixture of multivariate normals

$$p(\mathbf{x}_i) = \textstyle\sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{6}$$

and the joint probability of y_i and $\mathbf{x}_i$ is given by

$$p(y_i, \mathbf{x}_i; \Theta) = \textstyle\sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \times \int_{\mathbf{T}_{y_i}\mathbf{z}_{ik} \leq \mathbf{0}} \mathcal{N}(\mathbf{z}_{ik}; \mathbf{W}_k^T\mathbf{x}_i + \mathbf{b}_k, \mathbf{I}) d\mathbf{z}_{ik} \tag{7}$$

where $\Theta = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \mathbf{W}_k, \mathbf{b}_k\}_{k=1}^K$ collects the parameters of interest in QGME.

Assume the feature vector $\mathbf{x}_i$ is partially observed, we partition it into $\mathbf{x}_i = [\mathbf{x}_i^{o_i}; \mathbf{x}_i^{m_i}]$, where $\mathbf{x}_i^{o_i}$ is the subvector of observed features and $\mathbf{x}_i^{m_i}$ is the subvector of missing features. The o_i and m_i denotes the set of indices for observed and missing features, respectively.

From (7) follows the joint probability of $\mathbf{x}_i^{o_i}$ and y_i

$$p(y_i, \mathbf{x}_i^{o_i}; \Theta) = \textstyle\sum_{k=1}^K \int \int_{\mathbf{T}_{y_i}\mathbf{z}_{ik} \leq \mathbf{0}} \pi_k \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \times \mathcal{N}(\mathbf{z}_{ik}; \mathbf{W}_k^T\mathbf{x}_i + \mathbf{b}_k, \mathbf{I}) d\mathbf{z}_{ik} d\mathbf{x}_i^{m_i} \tag{8}$$

and the marginal probability of $\mathbf{x}_i^{o_i}$

$$p(\mathbf{x}_i^{o_i}; \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K) = \textstyle\sum_{k=1}^K \int \pi_k \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) d\mathbf{x}_i^{m_i} \tag{9}$$

The conditional probability $p(y_i|\mathbf{x}_i^{o_i}; \Theta)$ follows directly from (8) and (9).

The following lemma is useful in deriving the ML estimator for the QGME. The proof is given in the Appendix.

Lemma 1. *Let* $\mathbf{x}_i = [\mathbf{x}_i^{o_i}; \mathbf{x}_i^{m_i}]$, $\boldsymbol{\mu}_k = [\boldsymbol{\mu}_k^{o_i}; \boldsymbol{\mu}_k^{m_i}]$, *and* $\boldsymbol{\Sigma}_k = \begin{bmatrix} \boldsymbol{\Sigma}_k^{o_i o_i} & \boldsymbol{\Sigma}_k^{o_i m_i} \\ \boldsymbol{\Sigma}_k^{m_i o_i} & \boldsymbol{\Sigma}_k^{m_i m_i} \end{bmatrix}$, *then*

$$\begin{aligned} &\mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\mathcal{N}(\mathbf{z}_{ik}; \mathbf{W}_k^T\mathbf{x}_i + \mathbf{b}_k, \mathbf{I}) \\ &= \mathcal{N}(\mathbf{x}_i^{o_i}; \boldsymbol{\mu}_i^{o_i}, \boldsymbol{\Sigma}_k^{o_i o_i})\mathcal{N}(\mathbf{z}_{ik}; \boldsymbol{\tau}_{ik}, \mathbf{G}_{ik})\mathcal{N}(\mathbf{x}_i^{m_i}; \mathbf{c}_{ik}, \mathbf{D}_{ik}) \end{aligned} \tag{10}$$

where

$$\boldsymbol{\tau}_{ik} = \mathbf{b}_k + \mathbf{W}_k^T\boldsymbol{\mu}_k + \boldsymbol{\Gamma}_{ik}^T(\boldsymbol{\Sigma}_k^{o_i o_i})^{-1}(\mathbf{x}_i^{o_i} - \boldsymbol{\mu}_k^{o_i}) \tag{11}$$

$$\mathbf{G}_{ik} = \mathbf{I} + \mathbf{W}_k^T\boldsymbol{\Sigma}_k\mathbf{W}_k - \boldsymbol{\Gamma}_{ik}^T(\boldsymbol{\Sigma}_k^{o_i o_i})^{-1}\boldsymbol{\Gamma}_{ik} \tag{12}$$

$$\begin{aligned} \mathbf{c}_{ik} = {} & \boldsymbol{\mu}_k^{m_i} + \boldsymbol{\Sigma}_k^{m_i o_i}(\boldsymbol{\Sigma}_k^{o_i o_i})^{-1}(\mathbf{x}_i^{o_i} - \boldsymbol{\mu}_k^{o_i}) \\ & - \left[\boldsymbol{\Upsilon}_{ik} - \boldsymbol{\Sigma}_k^{m_i o_i}(\boldsymbol{\Sigma}_k^{o_i o_i})^{-1}\boldsymbol{\Gamma}_{ik}\right]\mathbf{G}_{ik}^{-1}\left[\mathbf{z}_{ik}\right. \\ & \left. - \mathbf{W}_k^T\boldsymbol{\mu}_k - \mathbf{b}_k - \boldsymbol{\Gamma}_{ik}^T(\boldsymbol{\Sigma}_k^{o_i o_i})^{-1}(\mathbf{x}_i^{o_i} - \boldsymbol{\mu}_k^{o_i})\right] \end{aligned} \tag{13}$$

$$\begin{aligned} \mathbf{D}_{ik} = {} & \boldsymbol{\Sigma}_k^{m_i m_i} - \boldsymbol{\Sigma}_k^{m_i o_i}\boldsymbol{\Sigma}_k^{o_i o_i\,-1}\boldsymbol{\Sigma}_k^{o_i m_i} \\ & -(\boldsymbol{\Upsilon}_{ik} - \boldsymbol{\Sigma}_k^{m_i o_i}\boldsymbol{\Sigma}_k^{o_i o_i\,-1}\boldsymbol{\Gamma}_{ik})\mathbf{G}_{ik}^{-1} \\ & \times(\boldsymbol{\Upsilon}_{ik} - \boldsymbol{\Sigma}_k^{m_i o_i}\boldsymbol{\Sigma}_k^{o_i o_i\,-1}\boldsymbol{\Gamma}_{ik})^T \end{aligned} \tag{14}$$

$$\boldsymbol{\Gamma}_{ik} = \boldsymbol{\Sigma}_k^{o_i o_i}\mathbf{W}_k^{o_i} + \boldsymbol{\Sigma}_k^{o_i m_i}\mathbf{W}_k^{m_i} \tag{15}$$

$$\boldsymbol{\Upsilon}_{ik} = \boldsymbol{\Sigma}_k^{m_i o_i}\mathbf{W}_k^{o_i} + \boldsymbol{\Sigma}_k^{m_i m_i}\mathbf{W}_k^{m_i} \tag{16}$$

3. ML Estimation of the QGME from Incomplete Data

The number of experts K is a hyper-parameter representing the complexity of QGME, which can be learned via cross-validation or Bayesian model selection. Given K, we estimate the parameters Θ through likelihood maximization.

For independent samples $\{(\mathbf{x}_i^{o_i}, y_i) : i = 1, \cdots, N\}$, the joint probability given by the QGME is $\prod_{i=1}^N p(y_i, \mathbf{x}_i^{o_i}; \Theta)$, taking logarithm of the joint probability, we get

$$\ell(\Theta) = \ln \textstyle\prod_{i=1}^N p(y_i, \mathbf{x}_i^{o_i}; \Theta) = \sum_{i=1}^N \ln p(y_i, \mathbf{x}_i^{o_i}; \Theta) \tag{17}$$

which is the likelihood function we wish to maximize. We choose to maximize the joint probability of $\mathbf{x}$'s and y's, instead of the conditional probability of y's given $\mathbf{x}$'s, because we are interested in both the classifier and the data manifold. As a matter of fact, the two are nonseparable given that the data are only partially observed. To find an accurate classifier, one needs to have accurate estimates of the missing features, which require a good estimate of the data manifold. In (Williams et al., 2005), the authors take a two-stage procedure: first the data manifold is estimated by maximizing the marginal probability of $\mathbf{x}$'s;

based on the estimated manifold, the missing feature in the classifier are integrated out to get a marginalized classifier, which is then used to estimate the parameters of the original classifier. Imputation methods are also based on such a two-stage procedure, with the marginalization implemented by sampling (finite imputation). The following theorem (the proof is given in the Appendix) shows that the two-stage procedure is not exact even when the integration over the missing feature is exact.

Theorem 2. *Let $p(\mathbf{x};\Theta_1)$ be the data manifold with parameter Θ_1, $p(y|\mathbf{x};\Theta_2)$ be the classifier with parameter Θ_2, and $\mathbf{x}=[\mathbf{x}^o;\mathbf{x}^m]$. Then $p(y|\mathbf{x}^o)$ depends on both Θ_1 and Θ_2.*

Since the marginalized classifier depends on both the parameter (Θ_1) of the data manifold and the parameter (Θ_2) of the original classifier, maximum likelihood estimation should be performed in the space of (Θ_1,Θ_2), instead of the space of Θ_1.

Because the manifold parameter and the classifier parameter are coupled in the presence of missing features[1], estimation of one requires knowing the other, which entails joint estimation of both. This motivates maximizing the joint likelihood function $\ell(\Theta)$ in (17).

Direct maximization of $\ell(\Theta)$ is difficult, since we must deal with the logarithm of sum and integrals. We apply Jensen's inequality to move the logarithm inside the sum and integrals, obtaining a lower bound of $\ell(\Theta)$ that is simpler to maximize. We first derive the lower bound of $\ln p(y_i,\mathbf{x}_i^{o_i};\Theta)$, corresponding to the i-th data sample; once that is done, the lower bound of $\ell(\Theta)$ is arrived by summing over data samples. It follows from (8) that

$$\begin{aligned}&\ln p(y_i,\mathbf{x}_i^{o_i};\Theta)\\&=\ln\sum_{k=1}^{K}\int_{\mathbf{T}_{y_i}\mathbf{z}_{ik}\le\mathbf{0}}\int q_{ik}(\mathbf{z}_{ik},\mathbf{x}_i^{m_i})\frac{\pi_k\mathcal{N}(\mathbf{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{q_{ik}(\mathbf{z}_{ik},\mathbf{x}_i^{m_i})}\\&\quad\times\mathcal{N}(\mathbf{z}_{ik};\mathbf{W}_k^T\mathbf{x}_i+\mathbf{b}_k,\mathbf{I})d\mathbf{x}_i^{m_i}d\mathbf{z}_{ik}\end{aligned}\tag{18}$$

where $\sum_{k=1}^{K}\int_{\mathbf{T}_{y_i}\mathbf{z}_{ik}\le\mathbf{0}}\int q_{ik}(\mathbf{z}_{ik},\mathbf{x}_i^{m_i})d\mathbf{x}_i^{m_i}d\mathbf{z}_{ik}=1$ and $q_{ik}(\mathbf{z}_{ik},\mathbf{x}_i^{m_i})\ge 0$. Applying Jensen's inequality to (18), we obtain

$$\begin{aligned}&\ln p(y_i,\mathbf{x}_i^{o_i};\Theta)\\&\ge\sum_{k=1}^{K}\int_{\mathbf{T}_{y_i}\mathbf{z}_{ik}\le\mathbf{0}}\int q_{ik}(\mathbf{z}_{ik},\mathbf{x}_i^{m_i})\ln\big[\pi_k\mathcal{N}(\mathbf{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\\&\quad\times\mathcal{N}(\mathbf{z}_{ik};\mathbf{W}_k^T\mathbf{x}_i+\mathbf{b}_k,\mathbf{I})\big]d\mathbf{x}_i^{m_i}d\mathbf{z}_{ik}\\&\quad+H[q_{ik}(\mathbf{z}_{ik},\mathbf{x}_i^{m_i})]\end{aligned}\tag{19}$$

where $H(\cdot)$ denotes Shannon entropy. The lower bound is tight (the equality holds) when

$$\begin{aligned}&q_{ik}(\mathbf{z}_{ik},\mathbf{x}_i^{m_i})\\&=\frac{\pi_k\mathcal{N}(\mathbf{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\mathcal{N}(\mathbf{z}_{ik};\mathbf{W}_k^T\mathbf{x}_i+\mathbf{b}_k,\mathbf{I})}{p(y_i,\mathbf{x}_i^{o_i};\Theta)}\end{aligned}\tag{20}$$

which, using Lemma 1, is rewritten as

$$\begin{aligned}&q_{ik}(\mathbf{z}_{ik},\mathbf{x}_i^{m_i})\\&=\delta_{ik}\mathcal{N}_{\mathbf{T}_{y_i}\mathbf{z}_{ik}\le 0}(\mathbf{z}_{ik};\boldsymbol{\tau}_{ik},\mathbf{G}_{ik})\mathcal{N}(\mathbf{x}_i^{m_i};\mathbf{c}_{ik},\mathbf{D}_{ik})\end{aligned}\tag{21}$$

where

$$\begin{aligned}&\mathcal{N}_{\mathbf{T}_{y_i}\mathbf{z}_{ik}\le 0}(\mathbf{z}_{ik};\boldsymbol{\tau}_{ik},\mathbf{G}_{ik})\\&=\begin{cases}\dfrac{\mathcal{N}(\mathbf{z}_{ik};\boldsymbol{\tau}_{ik},\mathbf{G}_{ik})}{\int_{\mathbf{T}_{y_i}\mathbf{z}_{ik}\le 0}\mathcal{N}(\mathbf{z}_{ik};\boldsymbol{\tau}_{ik},\mathbf{G}_{ik})d\mathbf{z}_{ik}}, & \text{if } \mathbf{T}_{y_i}\mathbf{z}_{ik}\le 0\\ 0, & \text{otherwise}\end{cases}\end{aligned}\tag{22}$$

is a truncated multivariate normal distribution,

$$\delta_{ik}=\frac{\mathcal{N}(\mathbf{x}_i^{o_i};\boldsymbol{\mu}_k^{o_i},\boldsymbol{\Sigma}_k^{o_io_i})\int_{\mathbf{T}_{y_i}\mathbf{z}_{ik}\le 0}\mathcal{N}(\mathbf{z}_{ik};\boldsymbol{\tau}_{ik},\mathbf{G}_{ik})d\mathbf{z}_{ik}}{p(y_i,\mathbf{x}_i^{o_i};\Theta)}\tag{23}$$

and the normalization is calculated by

$$\begin{aligned}p(y_i,\mathbf{x}_i^{o_i};\Theta)=&\sum_{k=1}^{K}\mathcal{N}(\mathbf{x}_i^{o_i};\boldsymbol{\mu}_k^{o_i},\boldsymbol{\Sigma}_k^{o_io_i})\\&\times\int_{\mathbf{T}_{y_i}\mathbf{z}_{ik}\le 0}\mathcal{N}(\mathbf{z}_{ik};\boldsymbol{\tau}_{ik},\mathbf{G}_{ik})d\mathbf{z}_{ik}\end{aligned}\tag{24}$$

Define

$$\begin{aligned}Q_i(\widehat{\Theta}|\Theta)=&\sum_{k=1}^{K}\int_{\mathbf{T}_{y_i}\mathbf{z}_{ik}\le\mathbf{0}}\int q_{ik}(\mathbf{z}_{ik},\mathbf{x}_i^{m_i})\\&\times\ln\big[\widehat{\pi}_k\mathcal{N}(\mathbf{x}_i;\widehat{\boldsymbol{\mu}}_k,\widehat{\boldsymbol{\Sigma}}_k)\\&\times\mathcal{N}(\mathbf{z}_{ik};\widehat{\mathbf{W}}_k^T\mathbf{x}_i+\widehat{\mathbf{b}}_k,\mathbf{I})\big]d\mathbf{x}_i^{m_i}d\mathbf{z}_{ik}\\&+H[q_{ik}(\mathbf{z}_{ik},\mathbf{x}_i^{m_i})]\end{aligned}\tag{25}$$

By (19), $Q_i(\widehat{\Theta}|\Theta)\le Q_i(\widehat{\Theta}|\widehat{\Theta})$. If we make $Q_i(\widehat{\Theta}|\Theta)\ge Q_i(\Theta|\Theta)$, we have $\ln p(y_i,\mathbf{x}_i^{o_i};\Theta)=Q_i(\Theta|\Theta)\le Q_i(\widehat{\Theta}|\widehat{\Theta})=\ln p(y_i,\mathbf{x}_i^{o_i};\widehat{\Theta})$. The inequality still holds if we sum over samples i. Thus, if we use (21) to compute $q_{ik}(\mathbf{z}_{ik},\mathbf{x}_i^{m_i})$ and let $\widehat{\Theta}=\arg\max_{\widehat{\Theta}}\sum_{i=1}^{N}Q_i(\widehat{\Theta}|\Theta)$, we are assured that $\widehat{\Theta}$ is improved over Θ. This gives the expectation-maximization (EM) algorithm for estimating Θ.

We now discuss calculation of $Q_i(\widehat{\Theta}|\widehat{\Theta})$. Introducing

$$\begin{aligned}\mathbf{r}_{ik}=&\boldsymbol{\mu}_k^{m_i}+\boldsymbol{\Sigma}_k^{m_io_i}(\boldsymbol{\Sigma}_k^{o_io_i})^{-1}(\mathbf{x}_i^{o_i}-\boldsymbol{\mu}_k^{o_i})\\&+\big[\boldsymbol{\Upsilon}_{ik}-\boldsymbol{\Sigma}_k^{m_io_i}(\boldsymbol{\Sigma}_k^{o_io_i})^{-1}\boldsymbol{\Gamma}_{ik}\big]\mathbf{G}_{ik}^{-1}[\mathbf{W}_k^T\boldsymbol{\mu}_k\\&+\mathbf{b}_k+\boldsymbol{\Gamma}_{ik}^T(\boldsymbol{\Sigma}_k^{o_io_i})^{-1}(\mathbf{x}_i^{o_i}-\boldsymbol{\mu}_k^{o_i})\big]\end{aligned}\tag{26}$$

$$\boldsymbol{\Omega}_{ik}=\big[\boldsymbol{\Upsilon}_{ik}-\boldsymbol{\Sigma}_k^{m_io_i}(\boldsymbol{\Sigma}_k^{o_io_i})^{-1}\boldsymbol{\Gamma}_{ik}\big]\mathbf{G}_{ik}^{-1}\tag{27}$$

we simplify (13) to

$$\mathbf{c}_{ik}=\mathbf{r}_{ik}+\boldsymbol{\Omega}_{ik}\mathbf{z}_{ik}\tag{28}$$

[1]The manifold and classifier parameters are coupled in QGME even when all features are observed, since the manifold is used to form the gating network, which is a part of the classifier. But the complete data case is not the focus in this paper.

Theorem 3. *Let E denote expectation with respect to the truncated normal in (22), and*

$$\boldsymbol{\beta}_{ik} = \mathbf{r}_{ik} + \boldsymbol{\Omega}_{ik} E(\mathbf{z}_{ik}) \tag{29}$$

$$\mathbf{F}_{ik} = \boldsymbol{\Omega}_{ik} E[(\mathbf{z}_{ik} - E\mathbf{z}_{ik})(\mathbf{z}_{ik} - E\mathbf{z}_{ik})^T]\boldsymbol{\Omega}_{ik}^T + \mathbf{D}_{ik} \tag{30}$$

where the expectation E is taken with respect to the truncated normal in (22). Then $Q_i(\widehat{\Theta}|\Theta) = \sum_{k=1}^K \delta_{ik}\Psi_i + H[q_{ik}(\mathbf{z}_{ik}, \mathbf{x}_i^{m_i})]$, with

$$\begin{aligned}\Psi_i = \ln & \frac{1}{(2\pi)^{\frac{d+1}{2}}|\widehat{\boldsymbol{\Sigma}}_k|^{\frac{1}{2}}} \\ & -\frac{1}{2}\left(\begin{bmatrix}\mathbf{x}_i^{o_i}\\ \boldsymbol{\beta}_{ik}\end{bmatrix} - \widehat{\boldsymbol{\mu}}_k\right)^T \widehat{\boldsymbol{\Sigma}}_k^{-1}\left(\begin{bmatrix}\mathbf{x}_i^{o_i}\\ \boldsymbol{\beta}_{ik}\end{bmatrix} - \widehat{\boldsymbol{\mu}}_k\right) \\ & -\frac{1}{2}\mathrm{trace}\left(\widehat{\boldsymbol{\Sigma}}_k^{-1}\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{F}_{ik}\end{bmatrix}\right) \\ & -\frac{1}{2}\left\|E(\mathbf{z}_{ik}) - \widehat{\mathbf{W}}_k^T\begin{bmatrix}\mathbf{x}_i^{o_i}\\ \boldsymbol{\beta}_{ik}\end{bmatrix} - \widehat{\mathbf{b}}_k\right\|^2 \\ & -\frac{1}{2}\mathrm{trace}\left(\widehat{\mathbf{W}}_k^T\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{F}_{ik}\end{bmatrix}\widehat{\mathbf{W}}_k\right) \\ & -\frac{1}{2}E[(\mathbf{z}_{ik} - E\mathbf{z}_{ik})(\mathbf{z}_{ik} - E\mathbf{z}_{ik})^T]\end{aligned} \tag{31}$$

The proof is given in the Appendix. Calculation of $\sum_{i=1}^N Q_i(\widehat{\Theta}|\Theta)$ completes the E-step, where the missing features $\mathbf{x}_i^{m_i}$ are replaced with the expected values $\boldsymbol{\beta}_{ik}$ in the k-th expert. Such imputations are adaptive since the expected values change as the QGME parameters are updated in the M-step. With δ_{ik} given by (23), the M-step is achieved by maximizing $\sum_{i=1}^N Q_i(\widehat{\Theta}|\Theta)$ to obtain

$$\widetilde{\delta}_{ik} = \frac{\delta_{ik}}{\sum_{i=1}^N \delta_{ik}}, \qquad \widehat{\boldsymbol{\mu}}_k = \sum_{i=1}^N \widetilde{\delta}_{ik}\begin{bmatrix}\mathbf{x}_i^{o_i}\\ \boldsymbol{\beta}_{ik}\end{bmatrix}$$

$$\widehat{\boldsymbol{\Sigma}}_k = \sum_{i=1}^N \widetilde{\delta}_{ik}\left\{\left(\begin{bmatrix}\mathbf{x}_i^{o_i}\\ \boldsymbol{\beta}_{ik}\end{bmatrix} - \widehat{\boldsymbol{\mu}}_k\right)\left(\begin{bmatrix}\mathbf{x}_i^{o_i}\\ \boldsymbol{\beta}_{ik}\end{bmatrix} - \widehat{\boldsymbol{\mu}}_k\right)^T + \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{F}_{ik}\end{bmatrix}\right\}$$

$$\begin{bmatrix}\widehat{\mathbf{b}}_k^T\\ \widehat{\mathbf{W}}_k\end{bmatrix} = \left(\sum_{i=1}^N \widetilde{\delta}_{ik}\left(\begin{bmatrix}1\\ \mathbf{x}_i^{o_i}\\ \boldsymbol{\beta}_{ik}\end{bmatrix}\begin{bmatrix}1\\ \mathbf{x}_i^{o_i}\\ \boldsymbol{\beta}_{ik}\end{bmatrix}^T + \begin{bmatrix}0 & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{F}_{ik}\end{bmatrix}\right)\right)^{-1} \times \sum_{i=1}^N \widetilde{\delta}_{ik}\begin{bmatrix}1\\ \mathbf{x}_i^{o_i}\\ \boldsymbol{\beta}_{ik}\end{bmatrix} E(\mathbf{z}_{ik})$$

Computation of the moments of $\mathbf{z}_{ik}$ involve integrals of (multivariate) normal distributions over the region specified by $\mathbf{T}_{y_i}\mathbf{z}_{ik} \leq \mathbf{0}$. In the binary case, $M = 2$ and z_{ik} is univariate truncated normal, with $T_0 = 1$ and $T_1 = -1$. In this case, we have $E(z_{ik}) = \tau_{ik}\Phi(\frac{\tau_{ik}}{\sqrt{G_{ik}}}) + \sqrt{G_{ik}}\mathcal{N}(\frac{\tau_{ik}}{\sqrt{G_{ik}}}; 0, 1)$ and $E(z_{ik}^2) = (\tau_{ik}^2 + G_{ik})\Phi(\frac{\tau_{ik}}{\sqrt{G_{ik}}}) + \tau_{ik}\sqrt{G_{ik}}\mathcal{N}(\frac{\tau_{ik}}{\sqrt{G_{ik}}}; 0, 1)$, where $\Phi(\cdot)$ is the cumulative distribution function (cdf) of the standard normal.

In the trinary case, $M = 3$ and z_{ik} is a bivariate truncated normal. One can use the formulae in (Rosenbaum, 1961) to compute the moments analytically. In the general case when $M > 3$, the results in (Genz, 1992) can be employed to perform the multiple integral numerically.

4. Experimental Results

We demonstrate the performance of QGME on three benchmark data sets — Johns Hopkins University Ionosphere database (Ionosphere), Wisconsin Diagnostic Breast Cancer (WDBC) data, and Iris Plant Database (Iris). These data sets and their descriptions are publicly available at the UCI machine learning repository (Newman et al., 1998).

4.1. Results on Ionosphere and WDBC

In the first experiment, we compare the QGME to the method in (Williams et al., 2005), which represents the most recent development in incomplete data classification. To replicate the results from previous experiments, we use the same data sets, i.e., Ionosphere and WDBC, and follow the same experimental procedure, as used in (Williams et al., 2005). Specifically, for each data sample, we randomly select a portion of features to be observed, assuming the remaining are missing. This simulates the missing completely at random (MCAR) case (Schafer & Graham, 2002). Note that the observed and missing features change from sample to sample, thus we are considering general missing patterns. For each of the two data sets, we randomly partition the data samples into training and testing subsets. We consider different experimental settings, by varying the percentage of missing features and the percentage of data samples used in training. For each experimental setting, we perform 100 independent trials, each consisting of a random feature partition and a random sample partition that are performed independently of each other. Following Williams et al., the area under ROC curve (AUC) (Hanley & McNeil, 1982) is used as algorithms' performance measure.

The results are summarized in Figure 3, where each curve is the AUC as a function of the fraction of data samples used in training, resulting from an average over the independent trials, and the error bars on QGME show the standard deviations. The number of experts K controls the model complexity of QGME. To examine how the choice of K affects the generalization of QGME, we have plotted in Figure 3 the results for various values of K.

It is noted from Figure 3 that QGME outperforms the

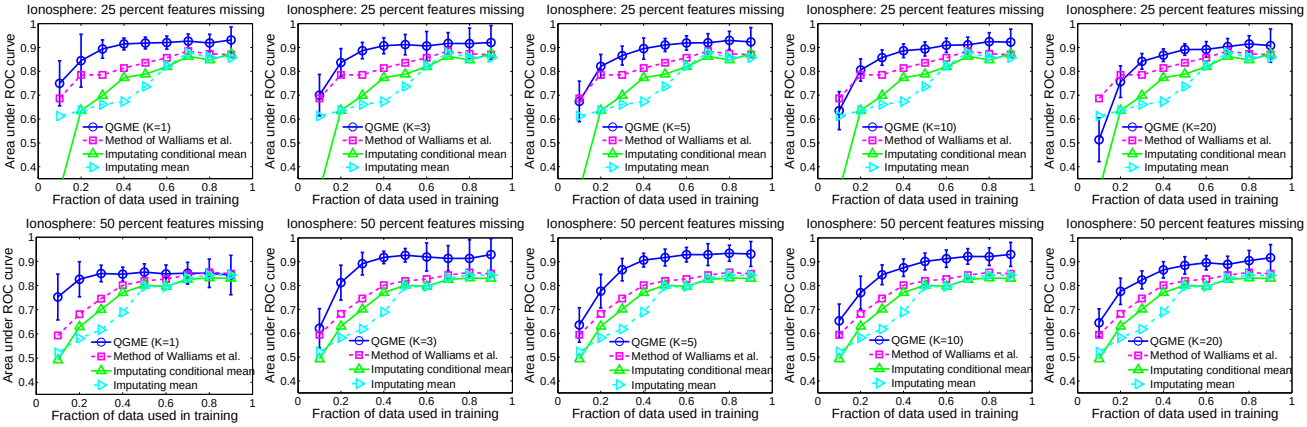

Figure 3. Results on Ionosphere data. Top: 25 percent features missing; Bottom: 50 percent features missing; From the leftmost column to the rightmost column, the number of experts used in QGME is $K = 1, 3, 5, 10, 20$. The results other than those of QGME are cited from (Williams et al., 2005).

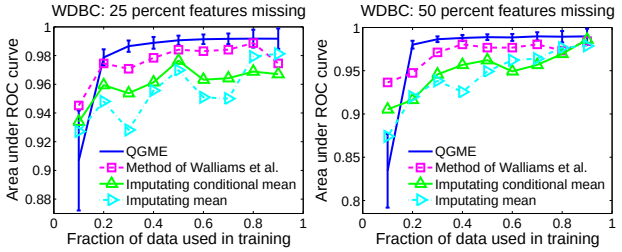

Figure 4. Results on WDBC data. Left: 25 percent features missing; Right: 50 percent features missing. The QGME implements a linear classifier by using $K = 1$. The results other than those of QGME are cited from (Williams et al., 2005).

competing methods, by significant margins, in almost all experimental settings. The improvements can be attributed to the adaptive analytic imputation accomplished inside the E-step of EM iterations. The adaptivity is a result of correctly handling the coupling of manifold and classifier parameters via joint estimation of both, as discussed in Section 3.

Within a wide range of choices for K (from 1 to 20), the QGME maintains its superiority in performance over the competing methods, which demonstrates its robustness to K. The relative insensitivity to K can be explained by the fact that the basic component (expert) in QGME is a linear classifier, the simplest of all classifiers. Since the expert is simple, the model complexity of QGME grows slow with the number of experts. Many real world data are linearly separable or nearly so, and a low complexity classifier like QGME generalizes better on these data.

Though the overall performance is robust to K, the choice of K does affect QGME in some cases, notably when the fraction of data used in training is less than 0.3. With few than 30% training data, a smaller K yields noticeably better performance than a larger K. This is understandable, considering that extremely insufficient training examples are highly prone to being overfitted, even by a model with low complexity.

A close observation of Figure 3 reveals that with more features missing (50%), QGME is less sensitive to K, showing that the internal adaptive imputation of missing features is robust to the choice of K, particularly when the data have heavily missing values. Since missing feature imputation is the key function for a method designed for incomplete data classification, this demonstrates the QGME to be a robust method for this purpose.

A last observation from Figure 3 is that the margin in performance improvement is larger when there are more features missing, which again shows the advantage of adaptive imputation in handling heavily missing values. In fact, imputation smooths out the data manifold and ameliorates generalization. The adaptive imputation takes into account the classifier when filling the missing values, and is thus particularly advantageous.

For the WDBC data, since they are easily separable data, we set $K = 1$ to make QGME a linear classifier. The results in Figure 4, where the error bars on QGME show the standard deviations, show that QGME outperforms the competing methods, except when there are fewer than 20% data used in training. The degradation shows that QGME is more easily affected by the number of training examples than the number of missing features. This also suggest that when training examples are very scarce, joint estimation of data manifold and classifiers may run into difficulties, since

it has more parameters to estimate and are more subject to over-fitting to outliers.

4.2. Results on a Three-class Problem

In the second experiment, we consider the Iris plant data, which is known as a three-class nonlinear problem. Instead of comparing to the method in (Williams et al., 2005), which can only deal with binary problems, we compare the QGME to support vector machines (Joachims, 1999), with the missing feature imputed with either mean or conditional mean. The conditional mean is obtained from the density $p(\mathbf{x})$ estimated by maximizing (6) with $K = 3$. The same value of K is used in QGME to make the results comparable. Estimation of $p(\mathbf{x})$ is performed using all incomplete data (labeled or unlabeled).

Since we have three classes, for which the AUC measure is not appropriate, the performance is evaluated in terms of correct classification rate, defined as the fraction of correctly classified data samples in the total number of data samples being tested. The results are summarized in Figure 5, which each curve is an average from 100 independent trials, each consisting a random feature partition and a random data partition, performed independently of each other. The error bars on QGME show the standard deviations.

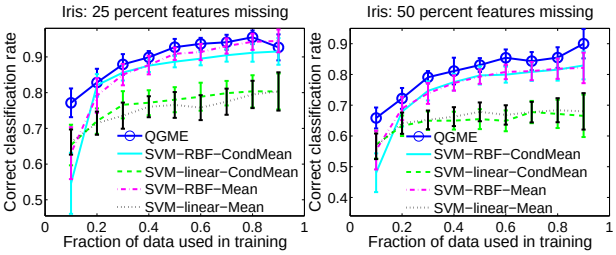

Figure 5. Results on Iris data. Left: 25 percent features missing; Right: 50 percent features missing. The QGME implements a nonlinear classifier by using $K = 3$. The imputation model for SVM is Eqn.(6) with $K = 3$, estimated using all incomplete data samples. The SVM results are generated by SVM$^{\text{light}}$, which is available at http://svmlight.joachims.org/. Two SVM classifiers are used: SVM-RBF uses a radial basis function (RBF) kernel and SVM-linear uses a linear kernel. The affix -Mean (or -CondMean) indicates the missing features are imputed with mean (or conditional mean).

It is seen from Figure 5 that QGME yields better performance than SVM, regardless of the kernel used and the imputation methods, and the improvement is over the entire range of experimental settings being investigated. The improvement is particularly prominent when there are more features missing (50%) and also when the training examples are few (10%).

Two more notes are made from Figure 5. First, SVM-RBF is much better than SVM-linear, demonstrating the nonlinearity of the problem. For either SVM-RBF or SVM-linear, imputing conditional means does not yields much improvement over imputing means, showing the limited utility of imputation when the imputation model and the classifier are learned in isolation. In contrast, QGME performs joint estimation of the two and is more advantageous.

5. Conclusions

We have proposed a statistical model, called *quadratically gated mixture of experts* (QGME), for multiclass nonlinear classification of incomplete data. The model uses linear classifiers as basic building blocks and mixes them through one-level quadratic gating. The model has an intrinsical design customized to data that are piecewise linearly separable. Many real world data fall under this category and can be analyzed by the model. The model handles missing values in a principled manner, via joint estimation of the imputation model and the classifier. Joint estimation solves the problem of parameter coupling between the data manifold, and the classifier and the resulting adaptive analytical imputation yields significant improvements over the methods that separate imputation model estimation from classifier learning, as demonstrated experimentally on three benchmark data sets.

Future work includes Bayesian model selection for K and fast computation of moments for latent variables z when there are more then three classes.

Appendix

Proof of Lemma 1

$$\mathcal{N}(\mathbf{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\mathcal{N}(\mathbf{z}_{ik};\mathbf{W}_k^T\mathbf{x}_i+\mathbf{b}_k,1) = \frac{\exp(-\frac{\psi_{ik}}{2})}{(2\pi)^{\frac{d+1}{2}}|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \quad \text{(A-1)}$$

where

$$\begin{aligned}\psi_{ik} &= (\mathbf{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_k)\\ &\quad +(\mathbf{z}_{ik}-\mathbf{W}_k^T\mathbf{x}_i-\mathbf{b}_k)^T(\mathbf{z}_{ik}-\mathbf{W}_k^T\mathbf{x}_i-\mathbf{b}_k)\\ &= \left(\begin{bmatrix}\mathbf{z}_{ik}\\ \mathbf{x}_i\end{bmatrix}-\begin{bmatrix}\mathbf{W}_k^T\boldsymbol{\mu}_k+\mathbf{b}_k\\ \boldsymbol{\mu}_k\end{bmatrix}\right)^T\\ &\quad\times\begin{bmatrix}\mathbf{I}+\mathbf{W}_k^T\boldsymbol{\Sigma}_k\mathbf{W}_k & \mathbf{W}_k^T\boldsymbol{\Sigma}_k\\ \boldsymbol{\Sigma}_k\mathbf{W}_k & \boldsymbol{\Sigma}_k\end{bmatrix}^{-1}\\ &\quad\times\left(\begin{bmatrix}\mathbf{z}_{ik}\\ \mathbf{x}_i\end{bmatrix}-\begin{bmatrix}\mathbf{W}_k^T\boldsymbol{\mu}_k+\mathbf{b}_k\\ \boldsymbol{\mu}_k\end{bmatrix}\right) \quad \text{(A-2)}\end{aligned}$$

Assuming partition of $\mathbf{x}_i$, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k$, as given in the premise, we have

$$\psi_{ik} = \left(\begin{bmatrix}\mathbf{z}_{ik}\\ \mathbf{x}_i^{o_i}\\ \mathbf{x}_i^{m_i}\end{bmatrix}-\begin{bmatrix}\mathbf{W}_k^T\boldsymbol{\mu}_k+\mathbf{b}_k\\ \boldsymbol{\mu}_i^{o_i}\\ \boldsymbol{\mu}_i^{m_i}\end{bmatrix}\right)^T$$

$$\times \begin{bmatrix} \mathbf{I}+\mathbf{W}_k^T\boldsymbol{\Sigma}_k\mathbf{W}_k & \boldsymbol{\Gamma}_{ik}^T & \boldsymbol{\Upsilon}_{ik}^T \\ \boldsymbol{\Gamma}_{ik} & \boldsymbol{\Sigma}_k^{o_i o_i} & \boldsymbol{\Sigma}_k^{o_i m_i} \\ \boldsymbol{\Upsilon}_{ik} & \boldsymbol{\Sigma}_k^{m_i o_i} & \boldsymbol{\Sigma}_k^{m_i m_i} \end{bmatrix}^{-1} \times \left(\begin{bmatrix} \mathbf{z}_{ik} \\ \mathbf{x}_i^{o_i} \\ \mathbf{x}_i^{m_i} \end{bmatrix} - \begin{bmatrix} \mathbf{W}_k^T\boldsymbol{\mu}_k+\mathbf{b}_k \\ \boldsymbol{\mu}_i^{o_i} \\ \boldsymbol{\mu}_i^{m_i} \end{bmatrix} \right) \quad \text{(A-3)}$$

where $\boldsymbol{\Gamma}_{ik}$ and $\boldsymbol{\Upsilon}_{ik}$ are as given in (15) and (16).

Since $\mathbf{z}_{ik}$, $\mathbf{x}_i^{o_i}$, and $\mathbf{x}_i^{m_i}$ are jointly normal distributed, $p(\mathbf{x}_i^{m_i}|\mathbf{z}_{ik},\mathbf{x}_i^{o_i})$ and $p(\mathbf{z}_{ik}|\mathbf{x}_i^{o_i})$ are also normal distributions. Then what remains to be proven is to verify that the mean and covariance matrix of $p(\mathbf{z}_{ik}|\mathbf{x}_i^{o_i})$ are given by (11) and (12), and those of $p(\mathbf{x}_i^{m_i}|\mathbf{z}_{ik},\mathbf{x}_i^{o_i})$ are given by (13) and (14), respectively. The verification if straightforward, using the properties of multivariate normal, and is thus omitted here. Q.E.D.

Proof of Theorem 2 It follows from the premise that

$$p(\mathbf{x}^o)=\int p(\mathbf{x};\Theta_1)d\mathbf{x}^m$$
$$p(y,\mathbf{x}^o)=\int p(y,\mathbf{x})d\mathbf{x}^m=\int p(y|\mathbf{x};\Theta_2)p(\mathbf{x};\Theta_1)d\mathbf{x}^m$$

Therefore, $p(y|\mathbf{x}^o) = \frac{p(y,\mathbf{x}^o)}{p(\mathbf{x}^o)} = \frac{\int p(y|\mathbf{x};\Theta_2)p(\mathbf{x};\Theta_1)d\mathbf{x}^m}{\int p(\mathbf{x};\Theta_1)d\mathbf{x}^m}$, which evidently depends on both Θ_1 and Θ_2. Q.E.D.

Proof of Theorem 3 Substituting (21) into (25)

$$Q_i(\widehat{\Theta}|\Theta)=\sum_{k=1}^{K}\delta_{ik}E_{\mathbf{z}_{ik}|\mathbf{x}_i^{o_i}}E_{\mathbf{x}_i^{m_i}|\mathbf{z}_{ik},\mathbf{x}_i^{o_i}}\Big\{\ln\big[\widehat{\pi}_k\mathcal{N}(\mathbf{x}_i;\widehat{\boldsymbol{\mu}}_k,\widehat{\boldsymbol{\Sigma}}_k) \times\mathcal{N}(\mathbf{z}_{ik};\widehat{\mathbf{W}}_k^T\mathbf{x}_i+\widehat{\mathbf{b}}_k,\mathbf{I})\big]\Big\}+H[q_{ik}(\mathbf{z}_{ik},\mathbf{x}_i^{m_i})] \quad \text{(A-4)}$$

where the expectation $E_{\mathbf{x}_i^{m_i}|\mathbf{z}_{ik},\mathbf{x}_i^{o_i}}$ is taken with respect to the normal $N(\mathbf{x}_i^{m_i};\mathbf{c}_{ik},\mathbf{D}_{ik})$ and $E_{\mathbf{z}_{ik}|\mathbf{x}_i^{o_i}}$ is taken with respect to the truncated normal in (22). The expectation inside the summation is evaluated as

$$\begin{aligned}
&E_{\mathbf{z}_{ik}|\mathbf{x}_i^{o_i}}E_{\mathbf{x}_i^{m_i}|\mathbf{z}_{ik},\mathbf{x}_i^{o_i}}\Big\{\ln\frac{1}{(2\pi)^{\frac{d+1}{2}}|\widehat{\boldsymbol{\Sigma}}_k|^{\frac{1}{2}}} \\
&-\frac{1}{2}(\mathbf{x}_i-\widehat{\boldsymbol{\mu}}_k)^T\widehat{\boldsymbol{\Sigma}}_k^{-1}(\mathbf{x}_i-\widehat{\boldsymbol{\mu}}_k)-\frac{1}{2}(\mathbf{z}_{ik}-(\widehat{\mathbf{W}}_k^T\mathbf{x}_i+\widehat{\mathbf{b}}_k))^2\Big\} \\
&=E_{\mathbf{z}_{ik}|\mathbf{x}_i^{o_i}}\Big\{\ln\frac{1}{(2\pi)^{\frac{d+1}{2}}|\widehat{\boldsymbol{\Sigma}}_k|^{\frac{1}{2}}}-\frac{1}{2}(\begin{bmatrix}\mathbf{x}_i^{o_i}\\ \mathbf{c}_{ik}\end{bmatrix}-\widehat{\boldsymbol{\mu}}_k)^T\widehat{\boldsymbol{\Sigma}}_k^{-1} \\
&\quad\times(\begin{bmatrix}\mathbf{x}_i^{o_i}\\ \mathbf{c}_{ik}\end{bmatrix}-\widehat{\boldsymbol{\mu}}_k)-\frac{1}{2}(\mathbf{z}_{ik}-(\widehat{\mathbf{W}}_k^T\begin{bmatrix}\mathbf{x}_i^{o_i}\\ \mathbf{c}_{ik}\end{bmatrix}+\widehat{\mathbf{b}}_k))^2 \\
&-\frac{1}{2}\mathrm{trace}(\mathbf{D}_{ik}\widehat{\boldsymbol{\Sigma}}_k^{-1\,m_i m_i})-\frac{1}{2}\mathrm{trace}(\widehat{\mathbf{W}}_k^{m_i\,T}\mathbf{D}_{ik}\widehat{\mathbf{W}}_k^{m_i})\Big\} \\
&=E_{\mathbf{z}_{ik}|\mathbf{x}_i^{o_i}}\Big\{\ln\frac{1}{(2\pi)^{\frac{d+1}{2}}|\widehat{\boldsymbol{\Sigma}}_k|^{\frac{1}{2}}} \\
&-\frac{1}{2}(\begin{bmatrix}\mathbf{0}\\ \boldsymbol{\Omega}_{ik}\mathbf{z}_{ik}\end{bmatrix}+\begin{bmatrix}\mathbf{x}_i^{o_i}\\ \mathbf{r}_{ik}\end{bmatrix}-\widehat{\boldsymbol{\mu}}_k)^T\widehat{\boldsymbol{\Sigma}}_k^{-1}(\begin{bmatrix}\mathbf{0}\\ \boldsymbol{\Omega}_{ik}\mathbf{z}_{ik}\end{bmatrix}+\begin{bmatrix}\mathbf{x}_i^{o_i}\\ \mathbf{r}_{ik}\end{bmatrix}-\widehat{\boldsymbol{\mu}}_k) \\
&-\frac{1}{2}\Big\|\mathbf{z}_{ik}-\widehat{\mathbf{W}}_k^T\begin{bmatrix}\mathbf{0}\\ \boldsymbol{\Omega}_{ik}\mathbf{z}_{ik}\end{bmatrix}-\widehat{\mathbf{W}}_k^T\begin{bmatrix}\mathbf{x}_i^{o_i}\\ \mathbf{r}_{ik}\end{bmatrix}-\widehat{\mathbf{b}}_k\Big\|^2 \\
&-\frac{1}{2}\mathrm{trace}(\mathbf{D}_{ik}\widehat{\boldsymbol{\Sigma}}_k^{-1\,m_i m_i})-\frac{1}{2}\widehat{\mathbf{W}}_k^{m_i\,T}\mathbf{D}_{ik}\widehat{\mathbf{W}}_k^{m_i}\Big\} \\
&=\ln\frac{1}{(2\pi)^{\frac{d+1}{2}}|\widehat{\boldsymbol{\Sigma}}_k|^{\frac{1}{2}}}-\frac{1}{2}(\begin{bmatrix}\mathbf{x}_i^{o_i}\\ \boldsymbol{\beta}_{ik}\end{bmatrix}-\widehat{\boldsymbol{\mu}}_k)^T\widehat{\boldsymbol{\Sigma}}_k^{-1}(\begin{bmatrix}\mathbf{x}_i^{o_i}\\ \boldsymbol{\beta}_{ik}\end{bmatrix}-\widehat{\boldsymbol{\mu}}_k) \\
&-\frac{1}{2}\mathrm{trace}(\widehat{\boldsymbol{\Sigma}}_k^{-1}\begin{bmatrix}\mathbf{0}&\mathbf{0}\\ \mathbf{0}&\mathbf{F}_{ik}\end{bmatrix})-\frac{1}{2}\Big\|E(\mathbf{z}_{ik})-\widehat{\mathbf{W}}_k^T\begin{bmatrix}\mathbf{x}_i^{o_i}\\ \boldsymbol{\beta}_{ik}\end{bmatrix}-\widehat{\mathbf{b}}_k\Big\|^2 \\
&-\frac{1}{2}\mathrm{trace}(\widehat{\mathbf{W}}_k^T\begin{bmatrix}\mathbf{0}&\mathbf{0}\\ \mathbf{0}&\mathbf{F}_{ik}\end{bmatrix}\widehat{\mathbf{W}}_k)-\frac{1}{2}E[(\mathbf{z}_{ik}-E\mathbf{z}_{ik})(\mathbf{z}_{ik}-E\mathbf{z}_{ik})^T]
\end{aligned}$$

where $\boldsymbol{\beta}_{ik}$ and $\mathbf{F}_{ik}$ are as given in the premise. Q.E.D.

References

Albert, J. H., & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, *88*, 669–679.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 141–149.

Ghahramani, Z., & Jordan, M. (1994). Supervised learning from incomplete data via the EM approach. *Advances in Neural Information Processing Systems*.

Hanley, J., & McNeil, B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29–36.

Ibrahim, J. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, *85*, 765–769.

Joachims, T. (1999). Making large-scale SVM learning practical. *Advances in Kernel Methods - Support Vector Learning*. MIT-Press.

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

Li, H., Liao, X., & Carin, L. (2006). Region-based value iteration for partially observable Markov decision processes. *The 23rd International Conference on Machine Learning (ICML)*.

Newman, D., Hettich, S., Blake, C., & Merz, C. (1998). UCI repository of machine learning databases. *http://www.ics.uci.edu/∼mlearn/MLRepository.html*.

Rosenbaum, S. (1961). Moments of a truncated bivariate normal distribution. *Journal of the Royal Statistical Society. Series B (Methodological)*, 405–408.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147–177.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*, 520–525.

Williams, D., Liao, X., Xue, Y., & Carin, L. (2005). Incomplete-data classification using logistic regression. *Proceedings of the 22nd International Machine Learning Conference*. ACM Press.