International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:9, No:12, 2015

Multinomial Dirichlet Gaussian Process Model for Classification of Multidimensional Data

Wanhyun Cho, Soonja Kang, Sangkyoon Kim, Soonyoung Park

***Abstract*—**We present probabilistic multinomial Dirichlet classification model for multidimensional data and Gaussian process priors. Here, we have considered efficient computational method that can be used to obtain the approximate posteriors for latent variables and parameters needed to define the multiclass Gaussian process classification model. We first investigated the process of inducing a posterior distribution for various parameters and latent function by using the variational Bayesian approximations and important sampling method, and next we derived a predictive distribution of latent function needed to classify new samples. The proposed model is applied to classify the synthetic multivariate dataset in order to verify the performance of our model. Experiment result shows that our model is more accurate than the other approximation methods.

***Keywords*—**Multinomial dirichlet classification model, Gaussian process priors, variational Bayesian approximation, Importance sampling, approximate posterior distribution, Marginal likelihood evidence.

I. Introduction

GAUSSIAN PROCESSES (GPs) are natural generations of multivariate Gaussian random variables to infinite index sets. Gaussian process is also a process that is constructed from classical statistical models by replacing latent functions of parametric form by random processes with Gaussian prior. GPs have been applied in a large number of fields in machine learning. One of these fields is the Gaussian regression and classification problem. In the case of regression with Gaussian noise, inference can be done simply in closed from, since the posterior is also a GP. However, in the case of classification, exact inference is analytically intractable because non-Gaussian likelihood function is assumed [1]-[3].

One prolific line of attack is based on approximating the non-Gaussian posterior with a tractable Gaussian distribution [4]-[6]. The most popular technique for treating intractable probabilistic models in these areas is the variational Bayesian approximation [7]. Thus, method approximated an intractable probability distribution by the closest distribution within a tractable family, where closeness is defined by the Kullback - Leibler divergence. In most applications, the tractable families contain distributions which factorize in all or in tractable subgroups of random variables. Hence, the method neglects correlations between variables which may be crucial in the learning of the hyper-parameters. However, despite these problems, empirical comparisons with exact analysis via MCMC and Laplace approximations illustrate the utility of the variational approximation as a computationally economic alternative to full MCMC and it is shown to be more accurate than the Laplace approximation.

Here, we will demonstrate that the variational Bayesian multinomial Dirichlet Gaussian process approach is efficient to classify multivariate data with Gaussian process priors. Our results have also motivated by the inclusion of the variational Gaussian approach within a larger study of different methods for classification with Gaussian processes.

Wanhyun Cho (professor) is with the Chonnam National University, Gwangju, 61186 South Korea (corresponding author, phone: +82-62-530-3443; fax +82-62-530-3449; e-mail whcho@ chnnam.ac.kr).

Soonja Kang (professor) is with the Chonnam National University, Gwangju, and 61186 South Korea (e-mail: sjkang@chonnam.ac.kr).

Sangkyoon Kim (researcher) and Soonyoung Park (professor) are with the Electrical Engineering Department, Mokpo National University, Chonnam, 58554 South Korea (e-mail: narciss76@moipo.ac.kr, sypark@mokpo.ac.kr).

II. Gaussian Process Classification Model

A. Our Model

Here, we are going to consider the multinomial Dirichlet Gaussian Process Classification Model (MDGPCM) defined such as. First, we have the data matrix as $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_N)^T$ which has dimension $(N \times D)$. Here, we represent the $(N \times K)$ dimensional matrix of Gaussian process latent variables as $\mathbf{F}$. And we also represent the $(N \times 1)$ dimensional columns of $\mathbf{F}$ and the $(K \times 1)$ dimensional rows of $\mathbf{F}$ as $\mathbf{f}^k, k = 1, \cdots, K$ and $\mathbf{f}_n, n = 1, \cdots, N$ respectively. Then, we assume that a GP prior for the k-th latent vector function $\mathbf{f}^k(\mathbf{x})$ is defined as the Gaussian distribution with zero mean vector and the $(N \times N)$ dimensional covariance matrix $\mathbf{K}^k$. That is, $\mathbf{f}^k \sim G(\mathbf{0}, \mathbf{K}^k)$. Here, the (i, j)-th element k_{ij}^k of the covariance matrix $\mathbf{K}^k$ will be defined by:

$$\mathbf{K}(\mathbf{X}, \boldsymbol{\varphi}^k) = (k_{ij}^k),\ k_{ij}^k = \exp\left(-\frac{1}{2}\sum_{d=1}^{D} \varphi_d^k (x_{id} - x_{jd})^2\right) \quad i, j = 1, \cdots, n \tag{1}$$

The $(D \times 1)$ vector of covariance kernel parameters for each class is denoted by $\boldsymbol{\varphi}^k$ and associated hyper-parameters $\boldsymbol{\psi}^k$ and hyper-hyper parameters $(\boldsymbol{\tau}^k, \zeta^k)$ complete the model.

Moreover, we define the $(NK \times 1)$ dimensional vector $\mathbf{f}(\mathbf{x})$ of the latent variables for K classes classification as:

$$\mathbf{f}(\mathbf{x}) = (\mathbf{f}^1, \cdots, \mathbf{f}^k, \cdots, \mathbf{f}^K)^T,$$
$$\mathbf{f}^k = (\mathrm{f}_1^k, \cdots, \mathrm{f}_n^k, \cdots \mathrm{f}_N^k)^T, k = 1, \cdots, K, \tag{2}$$

where the superscript k denotes a particular class and the subscript n denotes the observation number. Then, the GP prior of latent vector $\mathbf{f}(\mathbf{x})$ for K classes classification has usually been chosen to have only intra-class correlations. The covariance matrix for the prior of latent function $\mathbf{f}(\mathbf{x})$ is defined as:

$$\mathbf{K} = \begin{pmatrix} \mathbf{K}(\mathbf{X},\boldsymbol{\varphi}^1) & \cdots & \mathbf{0} \\ \vdots & \mathbf{K}(\mathbf{X},\boldsymbol{\varphi}^k) & \vdots \\ \mathbf{0} & \cdots & \mathbf{K}(\mathbf{X},\boldsymbol{\varphi}^K) \end{pmatrix} \tag{3}$$

Second, we define the $(N \times K)$ dimensional response matrix of associated target values as $\mathbf{Y} = (\mathbf{y}_1, \cdots, \mathbf{y}_N)^T$, where each component y_n^k of $(K \times 1)$ dimensional row vector $\mathbf{y}_n = (y_n^1, \cdots, y_n^K)^T$ takes zero or one, and satisfies the condition $\sum_{k=1}^{K} y_n^k = 1$ for all n. And we assume that a $(K \times 1)$ dimensional row vector $\mathbf{y}_n = (y_n^1, \cdots, y_n^K)^T$ is distributed with one observation multinomial distribution with parameter vector $\boldsymbol{\pi}_n = (\pi_n^1, \cdots, \pi_n^K)^T$. That is, the $\pi_n^k, k = 1, \cdots, K$ denotes the probability that the n-th observation $\mathbf{x}_n$ belongs to the k-th class. Moreover, we define the all of auxiliary random variables $\pi_n^k, k = 1, \cdots, K, n = 1, \cdots, N$ as a $(N \times K)$ dimensional random matrix $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \cdots, \boldsymbol{\pi}_N)^T$ with same size of response matrix $\mathbf{Y}$. We also assume that the prior distribution of n-th parameter vector $\boldsymbol{\pi}_n = (\pi_n^1, \cdots, \pi_n^K)^T$ of the $(N \times K)$ dimensional random matrix $\boldsymbol{\pi} = (\boldsymbol{\pi}_1, \cdots, \boldsymbol{\pi}_N)^T$ is given as the Dirichlet distribution with K parameter vector $\boldsymbol{\alpha}_n = (\alpha_n^1, \cdots, \alpha_n^K)$.

Third, we have to consider the link function that specifies the relation between the latent function $\mathbf{f}(\mathbf{x})$ and the response mean vector $E(\mathbf{Y} \mid \mathbf{f})$. To drive this relationship, we have first considered the $(N \times K)$ dimensional random matrix $\mathbf{M}$ of another auxiliary random variables $m_n^k, k = 1, \cdots, K, n = 1, \cdots, N$. Here, we have also assumed that each auxiliary random variable m_n^k is distributed with Gaussian $N(\mathrm{f}_n^k, 1)$. And we define K-dimensional parameter vector $\boldsymbol{\alpha}_n = (\alpha_n^1, \cdots, \alpha_n^K)$ of Dirichlet distribution $Dir(\boldsymbol{\pi}_n \mid \boldsymbol{\alpha}_n)$ as K parameters $\boldsymbol{\alpha}_n = (\exp(m_n^1), \cdots, \exp(m_n^K))$. Then, the relationship between the response row vector $\mathbf{y}_n = (y_n^1, \cdots, y_N^K)^T$ and the auxiliary latent vector $\boldsymbol{\pi}_n = (\pi_n^1, \cdots, \pi_n^K)^T$ is adopted as the following form:

$$y_n^k = 1 \text{ if } \pi_n^k = \max_{1 \le j \le K} \{\pi_n^j\} \tag{4}$$

Moreover, by using an expectation property of Dirichlet distribution and the softmax method, the probability that a response variable y_n^k takes 1 can be defined by:

$$p(y_n^k = 1 \mid \mathbf{m}_n) = E(\pi_n^k \mid \boldsymbol{\alpha}_n = \exp(\mathbf{m}_n)) = \frac{\exp(m_n^k)}{\sum_{l=1}^{K} \exp(m_n^l)} \tag{5}$$
$$k = 1, \cdots, K,\ n = 1, \cdots, N.$$

Therefore, since the random variables m_n^k is distributed with the Gaussian distribution $N(\mathrm{f}_n^k, 1)$, the relation between the probability $p(y_n^k = 1 \mid \mathbf{m}_n)$ and latent function $\mathbf{f}_n$ can be approximated as:

$$p(y_n^k = 1 \mid \mathbf{m}_n) = \frac{\exp(m_n^k)}{\sum_{l=1}^{K} \exp(m_n^l)} \approx \frac{\exp(\mathrm{f}_n^k)}{\sum_{l=1}^{K} \exp(\mathrm{f}_n^l)} \tag{6}$$
$$k = 1, \cdots, K,\ n = 1, \cdots, N.$$

Furthermore, an hierarchic prior on the covariance function parameters $\boldsymbol{\varphi}^k$ is employed such that each parameter has an independent exponential distribution $\varphi_d^k \sim Exp(\psi_d^k)$, $d = 1, \cdots, D$ and a gamma distribution is placed on the mean value of the exponential $\psi_d^k \sim \Gamma(\sigma^k, \tau^k), d = 1, \cdots, D$. Thus, they are forming a conjugate pair. The associated hyper-hyper-parameters $\boldsymbol{\omega} = (\sigma^{k=1,\cdots,K}, \tau^{k=1,\cdots,K})$ can be set to reflect some prior knowledge of the data. Alternatively, vague priors can be employed such that, for example, each $\sigma^{k=1,\cdots,K} = \tau^{k=1,\cdots,K} = 1$.

Finally, if we define a set of the hidden variables as $\Theta = \{\boldsymbol{\pi}, \mathbf{m}, \mathbf{f}\}$ and the parameters as $\Phi = \{\boldsymbol{\varphi}^{k=1,\cdots,K}, \boldsymbol{\psi}^{k=1,\cdots,K}\}$, the joint likelihood function for all hidden variables and parameters takes the following form:

$$\begin{aligned} p(\mathbf{Y}, \Theta, \Phi \mid \mathbf{X}, \boldsymbol{\omega}) = & \prod_{n=1}^{N} p(\mathbf{y}_n \mid \boldsymbol{\pi}_n) p(\boldsymbol{\pi}_n \mid \boldsymbol{\alpha}_n = \exp(\mathbf{m}_n)) \\ & \times \prod_{k=1}^{K} p(\mathbf{m}^k \mid \mathbf{f}^k) p(\mathbf{f}^k \mid \mathrm{K}_{(N \times N)}^k(\mathbf{X}, \boldsymbol{\varphi}^k)) \\ & \times p(\boldsymbol{\varphi}^k \mid \boldsymbol{\psi}^k) p(\boldsymbol{\psi}^k \mid \boldsymbol{\omega}^k) \end{aligned} \tag{7}$$

where the individual factors are respectively define as: $p(\mathbf{Y} \mid \boldsymbol{\pi})$ is one observation multinomial distribution with probability vector $\boldsymbol{\pi}_n$, $p(\boldsymbol{\pi} \mid \boldsymbol{\alpha})$ is the Dirichlet distribution with parameter vector $\boldsymbol{\alpha}_n = \exp(\mathbf{m}_n)$, $p(\mathbf{m} \mid \mathbf{f})$ is the Gaussian distribution with mean f_n^k and variance 1, $p(\mathbf{f} \mid \mathbf{X}, \boldsymbol{\varphi})$ is a multivariate Gaussian distribution with zero mean vector and covariance matrix $\mathrm{K}_{(n \times n)}^k(\mathbf{X}, \boldsymbol{\varphi}^k)$, $p(\boldsymbol{\varphi} \mid \boldsymbol{\psi})$ is an exponential distribution with parameter ψ_d^k, and . $p(\boldsymbol{\psi} \mid \boldsymbol{\omega})$ is the gamma distribution with parameter (σ^k, τ^k).

B. The Variational Bayesian Approximation

We now consider an approximate variational Bayesian inference to drive the posterior distributions for latent function

and parameters in our model. This method is a particular variational method which aims to find some approximate joint distribution $q(\Theta) = q(\boldsymbol{\pi}, \mathbf{m}, \mathbf{f})$ over latent variables $\Theta = \{\boldsymbol{\pi}, \mathbf{m}, \mathbf{f}\}$ to approximate the true joint distribution $p(\Theta) = p(\boldsymbol{\pi}, \mathbf{m}, \mathbf{f})$ by minimizing the KL-divergence $KL(q(\Theta) \mid p(\Theta)$ defined as:

$$KL(q(\Theta) \mid p(\Theta \mid \mathbf{Y}, \mathbf{X}, \Theta)) = -\int q(\Theta) \ln\left(\frac{p(\Theta \mid \mathbf{Y}, \mathbf{X}, \Phi, \boldsymbol{\omega})}{q(\Theta)}\right) d\Theta \quad (8)$$

Then, vatiational distribution $q(\Theta)$ is taken by minimizing the KL divergence, and it is equivalent to maximizing the free energy. By the way, the free energy is divided into two components as:

$$F(q, \Theta) = \int q(\Theta) \ln p(\mathbf{Y} \mid \mathbf{X}, \Phi, \boldsymbol{\omega}) d\Theta - \int q(\Theta) \ln q(\Theta) d\Theta \quad (9)$$

Here, the variational distribution $q(\Theta)$ is usually assumed to factorize over some partition of latent variables $\Theta = \{\boldsymbol{\pi}, \mathbf{m}, \mathbf{f}\}$ like as $q(\Theta) = q(\boldsymbol{\pi})q(\mathbf{m})q(\mathbf{f})$ **.** It can be shown using the calculus of variations that the best distribution $q^*(\Theta_i)$ for each of the factors $q(\Theta_i), i = 1,2,3$ (in term of the distribution maximizing the free energy) can be expressed as:

$$q^*(\Theta_i) \approx \exp(E_{q(\Theta \setminus \Theta_i)}(\ln p(\mathbf{Y}, \Theta \mid \mathbf{X}, \Phi, \boldsymbol{\omega}))), i = 1,2,3 \quad (10)$$

where $E_{q(\Theta \setminus \Theta_i)}(\ln p(\mathbf{Y}, \Theta \mid \mathbf{X}, \Phi, \boldsymbol{\alpha}))$ is the expectation of the logarithm of the joint probability of the data $\mathbf{Y}$ and latent variables Θ, taken over all variables not in the partition. In practice, we usually work in terms of logarithm of the variational distribution $q^*(\Theta_i)$, i.e.:

$$\ln q^*(\Theta_i) = E_{q(\Theta \setminus \Theta_i)}(\ln p(\mathbf{Y}, \Theta \mid \mathbf{X}, \Phi, \boldsymbol{\omega})) + \text{constant}, i = 1,2,3 \quad (11)$$

From a variational principle that we have considered so far, the variational Bayesian posterior distributions for latent function and all parameters can be summarized by the following iterations which, for all k and d, will optimize the bound on the marginal likelihood.

First, we consider the variational approximate posterior $q^*(\boldsymbol{\pi})$ for the parameter vector of classification probabilities $\boldsymbol{\pi}$.

$$q^*(\boldsymbol{\pi}) = \prod_{n=1}^{N} q^*(\boldsymbol{\pi}_n),\ q^*(\boldsymbol{\pi}_n) \sim \text{Dir}(\boldsymbol{\pi}_n \mid \boldsymbol{\beta}_n^*) \quad (12)$$

$\boldsymbol{\beta}_n^* = (\beta_n^{1*}, \cdots, \beta_n^{K*})$, where $\beta_n^{k*} = y_n^k + E_{q(m_n^k)}(\exp(\mathrm{m}_n^k))$.

Second, we consider the approximate posterior $q^*(\mathbf{m})$ of normal auxiliary variables $\mathbf{m}$ over Dirichlet auxiliary vector $\boldsymbol{\pi}$ and the latent variables $\mathbf{f}$. This is given as

$$q^*(\mathbf{m}) = \prod_{n=1}^{N}\prod_{k=1}^{K} q^*(m_n^k),\ q^*(m_n^k) = q_1^*(m_n^k) \times q_2^*(m_n^k) \quad (13)$$

where the distribution $q_1^*(m_n^k)$ is given by

$$q_1^*(m_n^k) = \frac{r(m_n^k)}{\sum_{l=1}^{K} r(m_n^l)}\ r(m_n^k) = \frac{\exp(\mathrm{E}_{q(\boldsymbol{\pi}_n)}[\ln \pi_n^k]\exp(m_n^k))}{\Gamma(\exp(m_n^k))} \quad (14)$$

and $\mathrm{E}_{q^*(\boldsymbol{\pi}_n)}[\ln \pi_n^k] = \upsilon(\beta_n^{k*}) - \upsilon(\sum_{l=1}^{K} \beta_n^{l*})$, $\upsilon(x)$ is the digamma function, and the distribution $q_2^*(m_n^k)$ is given by the normal distribution $N(m_n^k \mid \mathrm{E}_{q(\mathrm{f}_n^k)}(\mathrm{f}_n^k), 1)$. Therefore, the required posterior expectation $E(m_n^k)$ can be computed as the following manner by using importance sampling method:

$$E(m_n^k) = \int m_n^k q^*(m_n^k) dm_n^k \approx \sum_{l=1}^{L} m_n^k w(m_n^{k(l)}) \quad (15)$$

where each $m_n^{k(1)}, \cdots, m_n^{k(L)}$ are random samples drawn from;

$$N(m_n^k \mid \mathrm{E}_{q(\mathrm{f}_n^k)}(\mathrm{f}_n^k), 1), \text{ and } w(m_n^{k(l)}) = \frac{q_1^*(m_n^{k(l)})}{\sum_{s=1}^{L} q_1^*(m_n^{k(s)})} \quad (16)$$

Third, we consider the approximate posterior $q^*(\mathbf{f})$ of latent variables $\mathbf{f}$ over normal auxiliary variables $\mathbf{m}$ and the parameters $\boldsymbol{\varphi}^{k=1,\cdots,K}$. We have obtained the approximate posterior $q^*(\mathbf{f})$ for latent function $\mathbf{f}$:

$$q^*(\mathbf{f}) = \prod_{k=1}^{K} N\left(\mathbf{f}^k \mid \mathrm{E}(\mathbf{f}^k), \Sigma^k\right), \quad (17)$$

where $\mathrm{E}(\mathbf{f}^k) = \Sigma^k E(\mathbf{m}^k)$, and

$$\Sigma^k = \mathrm{K}_{(N \times N)}^k(\mathrm{E}_{q(\boldsymbol{\varphi}^k)}(\boldsymbol{\varphi}^k))(\mathbf{I} + \mathrm{K}_{(N \times N)}^k(\mathrm{E}_{q(\boldsymbol{\varphi}^k)}(\boldsymbol{\varphi}^k)))^{-1}$$

Fourth, if we also consider the set of hyper-parameters $\Phi = \{\boldsymbol{\varphi}^{k=1,\cdots,K}, \boldsymbol{\psi}^{k=1,\cdots,K}\}$, in this variational treatment, then the expectation of the covariance kernel hyper-parameters $\boldsymbol{\varphi}^{k=1,\cdots,K}$ under the variational posterior distribution $q(\boldsymbol{\varphi}^k)$ can be approximated by drawing S samples such that each $\varphi_s^{kd} \sim Exp(E_{q(\psi^{kd})}(\psi^{kd}))$ and so;

$$\mathrm{E}_{q(\boldsymbol{\varphi}^k)}[\boldsymbol{\varphi}^k] \approx \sum_{s=1}^{S} \boldsymbol{\varphi}_s^k w(\boldsymbol{\varphi}_s^k), \tag{18}$$

where,

$$w(\boldsymbol{\varphi}_s^k) = \frac{N(E_{q^*(\mathbf{f}^k)}(\mathbf{f}^k) \mid \mathbf{0}, \mathbf{K}(\mathbf{X}, \boldsymbol{\varphi}_s^k))}{\sum_{v=1}^{S} N(E_{q^*(\mathbf{f}^k)}(\mathbf{f}^k) \mid \mathbf{0}, \mathbf{K}(\mathbf{X}, \boldsymbol{\varphi}_v^k))}$$

Finally, the approximate posterior for hyper-parameters ψ_d^k is given as the gamma distribution with parameters $\sigma^k + 1$ and $\tau^k + \mathrm{E}_{q(\varphi_d^k)}(\varphi_d^k)$. And the required posterior mean values are given as;

$$E_{q(\psi_d^k)}(\psi_d^k) = \frac{\sigma^k + 1}{\tau^k + \mathrm{E}_{q(\varphi_d^k)}(\varphi_d^k)} \tag{19}$$

C. Predictive Distribution for New Observation

Here, we have first to consider a predictive distribution $q(\mathbf{f}_{\mathrm{new}} \mid \mathbf{X}, \mathbf{x}_*, \Theta, \mathbf{y})$ of latent function $\mathbf{f}_{new} = (\mathrm{f}_{new}^1, \cdots, \mathrm{f}_{new}^K)^T$ corresponding to new observation $\mathbf{x}_{new}$. Here, we define $\mathbf{K}^k(\mathbf{X} \mid \mathrm{E}(\varphi^k))$ as the $(N \times N)$ covariance matrix containing the covariance function values given by training data points $\mathbf{X}$, and we also define $\mathbf{k}_{new}^k(\mathbf{x}_{new}, \mathbf{X} \mid \mathrm{E}(\varphi^k))$ as the $(N \times 1)$ covariance vector containing the covariance function values between the new point $\mathbf{x}_{new}$ and those contained in training data $\mathbf{X}$. Moreover $k_{new}^k(\mathbf{x}_{new}, \mathbf{x}_{new} \mid \mathrm{E}(\varphi^k))$ denotes the covariance function value for the new point and itself.

Here, the predictive distribution $q(\mathbf{f}_{\mathrm{new}} \mid \mathbf{X}, \mathbf{x}_*, \Theta, \mathbf{y})$ will be Gaussian. Hence, its mean vector and covariance matrix are given as follows. The predictive mean vector for latent function $\mathbf{f}_{new}$ is given by:

$$\begin{aligned} \mathrm{E}_q(\mathbf{f}_{new} \mid \mathbf{X}, \mathbf{x}_*, \Theta, \mathbf{y}) &= Q(\mathbf{X}, \mathbf{x}_{new} \mid \mathrm{E}(\varphi))^T \mathbf{K}(\mathbf{X}, E(\boldsymbol{\varphi}))^{-1} \boldsymbol{\mu}_{\mathbf{f}} \\ &= Q(\mathbf{X}, \mathbf{x}_{new} \mid \mathrm{E}(\varphi))^T (\mathbf{I} + \mathbf{K}(\mathbf{X}, \mathrm{E}(\boldsymbol{\varphi})))^{-1} \mathrm{E}(\mathbf{m}) \end{aligned} \tag{20}$$

and the covariance matrix for latent function $\mathbf{f}_{new}$ is given by:

$$\begin{aligned} &Cov_q(\mathbf{f}_{new} \mid \mathbf{X}, \mathbf{y}, \Theta, \mathbf{x}_{\mathrm{new}}) = \mathrm{diag}(k_{new}^1, \cdots, k_{new}^K) \\ &-Q(\mathbf{X}, \mathbf{x}_{new} \mid \mathrm{E}(\varphi))^T (\mathbf{I} + \mathbf{K}(\mathbf{X}, E(\boldsymbol{\varphi})))^{-1} Q(\mathbf{X}, \mathbf{x}_{new} \mid \mathrm{E}(\varphi) \end{aligned} \tag{21}$$

When the mean $E(\mathrm{f}_{new}^k)$ and the variance $Var(\mathrm{f}_{new}^k)$ of latent function f_{new}^k for each class $k = 1, \cdots, K$ are given, we extract the S samples $\mathrm{f}_{new}^{k(1)}, \cdots, \mathrm{f}_{new}^{k(\mathrm{S})}$ of latent variable f_{new}^k from a normal distribution $N(\mathrm{f}_{new}^k \mid E(\mathrm{f}_{new}^k), Var(\mathrm{f}_{new}^k))$ having this mean and variance. And using a similar method, we also extract the S samples $m_{new}^{k(1)}, \cdots, m_{new}^{k(\mathrm{S})}$ of the auxiliary variables m_{new}^k for each class $k = 1, \cdots, K$ from the normal distribution $N(m_{new}^k \mid \mathrm{f}_{new}^k, 1)$ having a mean f_{new}^k and variance 1. Next, we define the parameters $\boldsymbol{\alpha}_{new}^s = (\exp(m_{new}^{1(\mathrm{s})}), \cdots, \exp(m_{new}^{K(\mathrm{s})}))$ of the Dirichlet probability distribution using the extracted samples $(m_{new}^{1(\mathrm{s})}, \cdots, m_{new}^{K(\mathrm{s})})$ and we extract again the classification probabilities vector $\boldsymbol{\pi}_{new}^{(s)} = (\pi_{new}^{1(\mathrm{s})}, \cdots, \pi_{new}^{K(\mathrm{s})})$ from the Dirichlet distribution.

$$\boldsymbol{\pi}_{new}^{(s)} \sim Dir((\pi_{new}^{1(\mathrm{s})}, \cdots, \pi_{new}^{K(\mathrm{s})}) \mid (\exp(m_{new}^{1(\mathrm{s})}), \cdots, \exp(m_{new}^{K(\mathrm{s})})). \tag{22}$$

Repeating this procedure S times, we have generated a total S number of classification probabilities vector $\boldsymbol{\pi}_{new}^{(1)}, \cdots, \boldsymbol{\pi}_{new}^{(\mathrm{S})}$. And using them, we calculate the mean of the classification probabilities vectors as:

$$\overline{\boldsymbol{\pi}}_{new} = \frac{1}{S}(\boldsymbol{\pi}_{new}^{(1)} + \cdots + \boldsymbol{\pi}_{new}^{(\mathrm{S})}) \tag{23}$$

III. Experimental Results

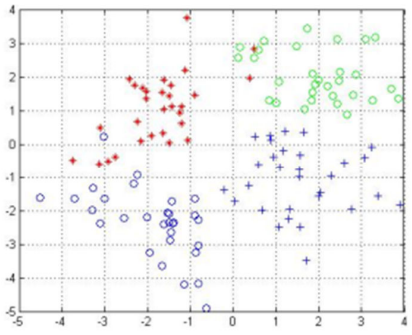
To estimate the performance of the proposed algorithm, we will consider four partially overlapping Gaussian sources of data in two dimensions. First, in order to train a model, we generate four classes' bivariate Gaussian random samples. One hundred twenty data points were generated by the four bivariate normal distributions with the mean vectors and covariance matrices described in Table I.

TABLE I
Mean Vector and Covariance Matrix for Each Class s

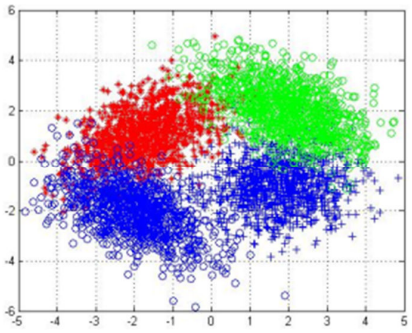
class	Mean vector	Covariance matrix
Class1	(1.75,-1.0)	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
Class2	(-1.75,1.0)	$\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$
Class3	(2,2)	$\begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$
Class4	(-2,-2)	$\begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$

One hundred and twenty draws were made from Table I and the test samples are used in the proposed algorithm with a further 4,000 points. Each of the sample values were sampled uniformly thus creating a balance of samples drawn from the four target classes. Fig. 1 (a) shows a plotting of training data points on a two-dimensional space. Second, in order to verify the performance of the model, we generate different four classes of bivariate Gaussian random sample. Four thousands data points were generated by the bivariate normal distribution.

Fig. 2 (a) shows the results that were monitored for covariance parameters during each step and as would be expected, it shows us a steady convergence in the improvements of the parameters. From Fig. 2 (b), we can also see that the development of the predictive performance follows that achieving a predictive performance of 92.2% at convergence. Therefore, it knows that the proposed method can entirely classify well data points.

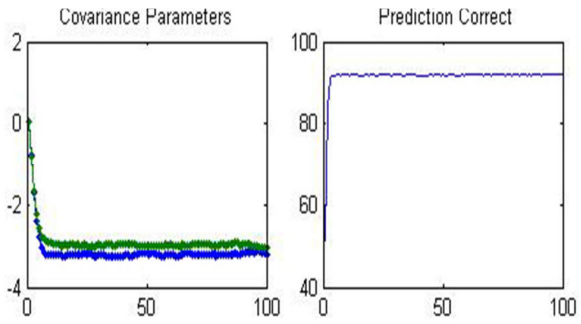


(a)



(b)

Fig. 1 (a) Training data, (b) Testing data



(a) (b)

Fig. 2 Experiments Results (a) Covariance Parameters φ_k, (b) Correct Prediction rate

IV. Conclusion

In this paper, we consider probabilistic multinomial Dirichlet classification model for multidimensional data with Gaussian process priors placed over the latent function. Variational Byesian algorithm is used to drive approximate posteriors for latent function as well as parameters needed to define the Multinomial dirichlect Gaussian process classification model. The proposed algorithm was performed in two steps: each of which is the training step and classification step. First, in the training step, using the variational Bayesian formula, we derived approximately the posterior distribution of the latent function and parameters on the basis of the learning data. Second, in the classification step, using a derived posterior distribution of the latent function, we estimate the classification probabilities to assign new sample into proper class. We assign this sample into a class that has the maximum probability.

Acknowledgment

This work was jointly supported by the National Research Foundation of Korea Government (2014R1A1A4A0109398) and the Research Foundation of Chonnam National University (2014-2256).

References

[1] C. E. Rasmussen, and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.
[2] H. Nicklisch, and C. E. Rasmussen, "Approximation for Binary Gaussian process Classification", *Journal of Machine Learning Research,* vol 9, pp 2035-75, 2008.
[3] C. K. I. Williams, and D. Barber, "Bayesian Classification with Gaussian Processes", *IEEE Tran. On PAMI*, vol 12, pp 1342-1351, 1998.
[4] T. P. Minka, "Expectation Propagation for Approximate Bayesian Inference", *In UAI, Morgan Kaufmann,* pp 362-369, 2001.
[5] M. Opper, and O. Winther, "Gaussian Processes for Classification: Mean Field Algorithms", *Neural Computation*, vol 12 pp 2655-2204, 2000.
[6] L. Csato, E. Fokoue, M. Opper, and B. Schottky, "Efficient Approaches to Gaussian Process Classification", *In Neural Information Processing Systems*, vol 12, pp 251-257, 2000.
[7] M. Girolami and S. Rogers, "Variational Bayesian Multinomial Probit Regression with Gaussian Process Priors", *Neural Computation*, vol 18 pp 1790-1817, 2006.

Wan-Hyun Cho received both B.S. degree and M.S. degree from the Department of Mathematics, Chonnam National University, Korea in 1977 and 1981, respectively and Ph.D. degree from the Department of Statistics, Korea University, Korea in 1988. He is now teaching in Chonnam National University. His research interests are statistical modeling, pattern recognition, image processing, and medical image processing.

Soon-Ja Kang received both B.S. degree and M.S. degree from the Department of Mathematics, Chonnam National University, Korea in 1979 and 1981, respectively and Ph.D. degree from the Department of Mathematics, Seogang University, Korea in 1988. She is now teaching in Chonnam National University. Her research fields are mathematical education, advanced calculus, and education for the gifted children.

Sang-Kyoon Kim received the B.S., M.S. and Ph.D. degrees in Electronics Engineering, Mokpo National University, Korea in 1998, 2000 and 2015 respectively. From 2011 to 2015, he was a Visiting Professor in the Department of Information & Electronics Engineering, Mokpo National University, Korea. His research interests include image processing, pattern recognition and computer vision.

Soon-Young Park received B.S. degree in Electronics Engineering from Yonsei University, Korea in 1982 and M.S and Ph.D. degrees in Electrical and Computer Engineering from State University of New York at Buffalo, in 1986 and 1989, respectively. From 1989 to 1990 he was a Postdoctoral Research Fellow in the department of Electrical and Computer Engineering at the State University of New York at Buffalo. Since 1990, he has been a Professor with Department of Electronics Engineering, Mokpo National University, Korea. His research interests include image and video processing, image protection and authentication and image retrieval techniques.