

# Interpretation of Hybrid Generative/Discriminative Algorithms

Jing-Hao Xue<sup>a,\*</sup>, D. Michael Titterington<sup>a</sup>

<sup>a</sup>*Department of Statistics, University of Glasgow, Glasgow G12 8QQ, UK*

---

## Abstract

In discriminant analysis, probabilistic generative and discriminative approaches represent two paradigms of statistical modelling and learning. In order to exploit the best of both worlds, hybrid modelling and learning techniques have attracted much research interest recently, one example being the so-called hybrid generative/discriminative algorithm proposed in Raina et al. (2003) and its multi-class extension (Fujino et al., 2007). In this paper, we interpret this hybrid algorithm from three perspectives, namely class-conditional probabilities, class-posterior probabilities and loss functions underlying the model. We suggest that the hybrid algorithm is by nature a generative model with its parameters learnt through both generative and discriminative approaches, in the sense that in fact it assumes a scaled data-generation process and uses scaled class-posterior probabilities to perform discrimination. Our suggestion can also be applied to the multi-class extension. In addition, using simulated data, we compare the performance of the normalised hybrid algorithm as a classifier with those of the naïve Bayes classifier and linear logistic regression. In general, our simulation studies suggest the following: if the covariance matrices are diagonal matrices, the naïve Bayes classifier performs the best; if the covariance matrices are non-diagonal matrices, linear logistic regression performs the best. In other words, our studies cannot support that the hybrid algorithm

offers better performance than either the naïve Bayes classifier or linear logistic regression alone, a phenomenon observed from the empirical studies in Raina et al. (2003)

*Key words:* Hybrid generative/discriminative models; Probabilistic generative and discriminative approaches; Statistical modelling and learning

---

## 1 Introduction

In recent years, under the new terminology of generative and discriminative approaches, the research interest in classical statistical modelling and learning approaches, namely the sampling paradigm and the diagnostic paradigm (Dawid, 1976; Titterington et al., 1981), to discriminant analysis has re-emerged in the machine learning community.

In discriminant analysis, observations with features  $\mathbf{x}$  measured are classified into classes labelled by a categorical variable  $y$ . The generative approach, such as normal-based discriminant analysis and the naïve Bayes classifier, models the joint distribution  $p(\mathbf{x}, y)$  of the features and the class labels factorised in the form  $p(\mathbf{x}|y)p(y)$ , and learns the model parameters through maximising the corresponding likelihood; the discriminative approach, such as logistic regression, models the conditional distribution  $p(y|\mathbf{x})$  of the class labels given the features, and learns the model parameters through maximising the corresponding conditional likelihood.

---

\* Corresponding author. Tel.: +44 141 330 2474; fax: +44 141 330 4814.

*Email addresses:* `jinghao@stats.gla.ac.uk` (Jing-Hao Xue),

`mike@stats.gla.ac.uk` (D. Michael Titterington).

1 Compared to the other, each of these two paradigms has its advantages and dis-  
2 advantages (Efron, 1975; Rubinstein and Hastie, 1997; Ng and Jordan, 2001).  
3 In order to exploit the best of both worlds, Bouchard and Triggs (2004) pro-  
4 pose the trade-off approach to modelling both  $p(y|\mathbf{x})$  and  $p(\mathbf{x}, y)$ , McCallum  
5 et al. (2006) propose the multi-conditional learning approach to modelling  
6  $p(y|\mathbf{x})$  and the data-generation process (DGP)  $p(\mathbf{x}|y)$ . In the sense that the  
7 generative and discriminative components within both approaches are derived  
8 from the joint distribution  $p(\mathbf{x}, y)$ , they can be regarded as hybrid learning of  
9 generative models.

10 Another interesting idea in this direction is the so-called hybrid generative/discriminative  
11 algorithm proposed by Raina et al. (2003), which introduces different weights  
12 to the partitions of the features within  $\mathbf{x}$ , learning most parameters genera-  
13 tively but the weights discriminatively. In this paper, we first interpret the  
14 hybrid algorithm from three perspectives, namely class-conditional probabili-  
15 ties, class-posterior probabilities and loss functions underlying the model, then  
16 discuss one of its multi-class extensions. Finally, by using simulated data, we  
17 compare its performance as a classifier with those of the naïve Bayes classifier  
18 and linear logistic regression.

## 19 **2 Interpretation of the Hybrid Algorithm**

20 Consider classifying an observation with  $h$  features into one of  $K$  groups by a  
21 classifier  $\hat{y}$ , which was trained by using the observed features and group labels  
22 of  $n$  other observations. We use an  $h$ -variate random vector  $\mathbf{x} = (x_1, \dots, x_h)^T$   
23 to represent the  $h$  features of the observation and a random categorical variable  
24  $y \in \{1, \dots, K\}$  to represent the group label. We denote a classifier of  $\mathbf{x}$  by

1  $\hat{y}(\mathbf{x})$  and the loss function of misclassifying  $\mathbf{x}$ , which arises from the group  $y$ ,  
 2 into the group  $\hat{y}(\mathbf{x})$  is  $L(y, \hat{y}(\mathbf{x}))$ .

### 3 2.1 Class-conditional Probabilities

4 For a binary classification where  $K = 2$ , based on Bayes' Theorem, the Bayes  
 5 discriminant criterion (*i.e.*,  $\hat{y}(\mathbf{x}) = \operatorname{argmax}_y p(y|\mathbf{x})$ ) of the generative classifiers  
 6 to classify  $\mathbf{x}$  into the group  $y = 1$  can be written as  $p(\mathbf{x}, y = 1) > p(\mathbf{x}, y = 2)$   
 7 or equivalently  $p(y = 1)p(\mathbf{x}|y = 1) \geq p(y = 2)p(\mathbf{x}|y = 2)$ . In addition, specific  
 8 generative classifiers, such as linear normal-based discriminant analysis with a  
 9 common diagonal covariance matrix (denoted by LDA- $\Lambda$ ) and the naïve Bayes  
 10 classifier, assume that the  $h$  features are conditionally independent given the  
 11 group label  $y$ , *i.e.*,  $p(\mathbf{x}|y) = \prod_{i=1}^h p(x_i|y)$ .

12 In the normalised hybrid and the unnormalised hybrid algorithms proposed  
 13 by Raina et al. (2003), the feature vector  $\mathbf{x}$  is divided into  $R$  partial feature  
 14 vectors  $\mathbf{x}^1, \dots, \mathbf{x}^R$ , because they suggest different levels of importance for  
 15 different partitions, or partial feature vectors; for example,  $\mathbf{x}^1$  may represent  
 16 the message subject of an email while  $\mathbf{x}^2$  represents the message body. As  
 17 with Raina et al. (2003), we focus on  $R = 2$  such that  $\mathbf{x} = (\mathbf{x}^{1T}, \mathbf{x}^{2T})^T$ ,  
 18  $\mathbf{x}^1 = (x_1, \dots, x_{h_1})^T$ ,  $\mathbf{x}^2 = (x_{h_1+1}, \dots, x_h)^T$  and  $h_2 = h - h_1$ , and assume that  
 19 the discriminant criterion of the generative classifiers can be rewritten as

$$p(y = 1)p(\mathbf{x}^1|y = 1)p(\mathbf{x}^2|y = 1) \geq p(y = 2)p(\mathbf{x}^1|y = 2)p(\mathbf{x}^2|y = 2) ;$$

20 and thus, given  $p(\mathbf{x}, y) \neq 0$ , the corresponding discriminant function  $\lambda_G(\mathbf{x}) =$

1  $\log \frac{p(y=1|\mathbf{x})}{p(y=2|\mathbf{x})}$  can be expressed in terms of likelihood ratios as

$$\lambda_G(\mathbf{x}) = \log \frac{p(y=1)}{p(y=2)} + \log \frac{p(\mathbf{x}^1|y=1)}{p(\mathbf{x}^1|y=2)} + \log \frac{p(\mathbf{x}^2|y=1)}{p(\mathbf{x}^2|y=2)} .$$

2 Such a representation can be obtained by assuming the generative DGP

$$p(\mathbf{x}|y) = w(\mathbf{x}^1, \mathbf{x}^2)p(\mathbf{x}^1|y)p(\mathbf{x}^2|y) ,$$

3 where  $w(\mathbf{x}^1, \mathbf{x}^2) \equiv 1$  such that, for all  $y$ ,  $\sum_{\mathbf{x}} p(\mathbf{x}|y) = 1$  with marginal distrib-  
 4 utions  $p(\mathbf{x}^1|y) = \sum_{\mathbf{x}^2} p(\mathbf{x}|y)$  and  $p(\mathbf{x}^2|y) = \sum_{\mathbf{x}^1} p(\mathbf{x}|y)$ , given that there exists  
 5  $\mathbf{x} = x$  such that  $p(x|y=1) \neq p(x|y=2)$ . In other words, it leads to assuming  
 6 conditional independence between partial feature vectors  $\mathbf{x}^1|y$  and  $\mathbf{x}^2|y$  such  
 7 that  $p(\mathbf{x}|y) = p(\mathbf{x}^1|y)p(\mathbf{x}^2|y)$ . To some extent, for a simple implementation in  
 8 practice, Raina et al. (2003) further assume that  $p(\mathbf{x}^1|y) = \prod_{j=1}^{h_1} p(x_j|y)$  and  
 9  $p(\mathbf{x}^2|y) = \prod_{j=h_1+1}^h p(x_j|y)$ ; these imply the conditional independence of the  
 10 elements within  $\mathbf{x}^1$  and  $\mathbf{x}^2$  given  $y$ , respectively.

11 Moreover, Raina et al. (2003) introduce two additional parameters  $\theta_1$  and  $\theta_2$   
 12 into the discriminant criterion, leading to different weights for different partial  
 13 feature vectors in the discrimination. Two ways of weighting are proposed  
 14 by Raina et al. (2003): one is

$$p(y=1)p(\mathbf{x}^1|y=1)^{\frac{\theta_1}{h_1}}p(\mathbf{x}^2|y=1)^{\frac{\theta_2}{h_2}} \geq p(y=2)p(\mathbf{x}^1|y=2)^{\frac{\theta_1}{h_1}}p(\mathbf{x}^2|y=2)^{\frac{\theta_2}{h_2}} ,$$

15 which is the criterion (denoted by *Criterion-H*) of the normalised hybrid al-  
 16 gorithm; the other is

$$p(y=1)p(\mathbf{x}^1|y=1)^{\theta_1}p(\mathbf{x}^2|y=1)^{\theta_2} \geq p(y=2)p(\mathbf{x}^1|y=2)^{\theta_1}p(\mathbf{x}^2|y=2)^{\theta_2} ,$$

17 which is the criterion of the unnormalised hybrid algorithm. Without loss of  
 18 generality, in this paper we focus on the normalised hybrid algorithm.

1 Let us write  $\theta = (\theta_1, \theta_2)^T$ . Then the hybrid algorithm can be derived from

$$p_\theta(\mathbf{x}|y) = w_\theta(\mathbf{x}^1, \mathbf{x}^2)p(\mathbf{x}^1|y)^{\frac{\theta_1}{h_1}}p(\mathbf{x}^2|y)^{\frac{\theta_2}{h_2}} \text{ and } p_\theta(\mathbf{x}, y) = p(y)p_\theta(\mathbf{x}|y) ,$$

2 where  $w_\theta(\mathbf{x}^1, \mathbf{x}^2)$  is independent of groups  $y$  so that it is cancelled out from  
 3 *Criterion-H*, but it is not necessarily further factorised in terms of  $w_\theta(\mathbf{x}^1, \mathbf{x}^2) =$   
 4  $w_\theta^1(\mathbf{x}^1)w_\theta^2(\mathbf{x}^2)$ . However, in order to maintain  $p_\theta(\mathbf{x}|y)$  as a proper probabil-  
 5 ity distribution (so that *Criterion-H* is derived from a proper probabilistic  
 6 model) with the marginal distributions  $p(\mathbf{x}^1|y) = \sum_{\mathbf{x}^2} p_\theta(\mathbf{x}|y)$  and  $p(\mathbf{x}^2|y) =$   
 7  $\sum_{\mathbf{x}^1} p_\theta(\mathbf{x}|y)$ , it is required, for all  $y$ , that

$$\sum_{\mathbf{x}^2} w_\theta(\mathbf{x}^1, \mathbf{x}^2)p(\mathbf{x}^2|y)^{\frac{\theta_2}{h_2}} = p(\mathbf{x}^1|y)^{1-\frac{\theta_1}{h_1}} , \text{ and}$$

8

$$\sum_{\mathbf{x}^1} w_\theta(\mathbf{x}^1, \mathbf{x}^2)p(\mathbf{x}^1|y)^{\frac{\theta_1}{h_1}} = p(\mathbf{x}^2|y)^{1-\frac{\theta_2}{h_2}} .$$

9 In some cases, it might be difficult to validate the existence of such a  $w_\theta(\mathbf{x}^1, \mathbf{x}^2)$ ,  
 10 *e.g.*, when  $\frac{\theta_1}{h_1} = 1$  while  $\frac{\theta_2}{h_2} \neq 1$  or vice versa, as the sums, in terms of  $\mathbf{x}$ , on the  
 11 left-hand sides of the above equations have to become independent of  $y$ . In  
 12 other cases, further assumptions might be needed to guarantee the existence.  
 13 We illustrate this by assuming that  $w_\theta(\mathbf{x}^1, \mathbf{x}^2)$  can be further factorised in  
 14 terms of  $w_\theta(\mathbf{x}^1, \mathbf{x}^2) = w_\theta^1(\mathbf{x}^1)w_\theta^2(\mathbf{x}^2)$ ; in other words, we assume conditional  
 15 independence between  $\mathbf{x}^1|y$  and  $\mathbf{x}^2|y$ . It follows that

$$p_\theta(\mathbf{x}|y) = w_\theta^1(\mathbf{x}^1)p(\mathbf{x}^1|y)^{\frac{\theta_1}{h_1}}w_\theta^2(\mathbf{x}^2)p(\mathbf{x}^2|y)^{\frac{\theta_2}{h_2}} ,$$

16 which also leads to *Criterion-H*. One solution of  $w_\theta(\mathbf{x}^1, \mathbf{x}^2)$  is, for all  $y$ ,

$$w_\theta^1(\mathbf{x}^1) = q(y)p(\mathbf{x}^1|y)^{1-\frac{\theta_1}{h_1}} , \quad w_\theta^2(\mathbf{x}^2) = \frac{1}{q(y)}p(\mathbf{x}^2|y)^{1-\frac{\theta_2}{h_2}} ,$$

17 where  $q(y)$  is a non-zero function used to cancel out terms of  $y$  within  $p(\mathbf{x}^1|y)^{1-\frac{\theta_1}{h_1}}$   
 18 and  $p(\mathbf{x}^2|y)^{1-\frac{\theta_2}{h_2}}$ . If such a  $w_\theta(\mathbf{x}^1, \mathbf{x}^2)$  cannot be found, *Criterion-H* is not a

1 Bayes discriminant criterion derived from a proper probabilistic model; nev-  
 2 ertheless, it is in practice still a discriminant criterion, although in this case  
 3 the hybrid algorithm is no longer a true Bayes classifier and, under a 0 – 1  
 4 loss function, it cannot provide a minimum Bayes error.

5 Under *Criterion-H*, we classify  $\mathbf{x}$  into  $y = 1$  if  $p_\theta(\mathbf{x}, y = 1) \geq p_\theta(\mathbf{x}, y = 2)$ .  
 6 Given  $p_\theta(\mathbf{x}, y) \neq 0$ , the discriminant function  $\lambda_H(\mathbf{x})$  of the hybrid algorithm  
 7 can be expressed in terms of weighted likelihood ratios as

$$\lambda_H(\mathbf{x}) = \log \frac{p(y = 1)}{p(y = 2)} + \frac{\theta_1}{h_1} \log \frac{p(\mathbf{x}^1|y = 1)}{p(\mathbf{x}^1|y = 2)} + \frac{\theta_2}{h_2} \log \frac{p(\mathbf{x}^2|y = 1)}{p(\mathbf{x}^2|y = 2)} .$$

8 Therefore,  $\lambda_H(\mathbf{x})$  can be viewed as a “weighted” version of the discriminant  
 9 function  $\lambda_G(\mathbf{x})$  of the generative classifier; however, as mentioned above, in  
 10 theory the hybrid algorithm should satisfy some conditions about the marginal  
 11 distributions to make the underlying model probabilistically valid. In addition,  
 12 as with  $\lambda_G(\mathbf{x})$ , most parameters in  $\lambda_H(\mathbf{x})$  are learnt by using a generative  
 13 approach; the weights are then learnt by using a discriminative approach based  
 14 on the learning results of the generative approach. Therefore, by nature the  
 15 hybrid algorithm can be regarded as a generative classifier since it assumes  
 16 the DGP  $p(\mathbf{x}|y)$  and thus  $p(\mathbf{x}, y)$ .

17 With the assumption of conditional independence between  $\mathbf{x}^1|y$  and  $\mathbf{x}^2|y$ , it  
 18 follows, between two class-conditional probabilities  $p(\mathbf{x}|y)$  and  $p_\theta(\mathbf{x}|y)$ , that

$$p_\theta(\mathbf{x}|y) = p(\mathbf{x}|y) \left\{ w_\theta(\mathbf{x}^1, \mathbf{x}^2) p(\mathbf{x}^1|y)^{\frac{\theta_1}{h_1}-1} p(\mathbf{x}^2|y)^{\frac{\theta_2}{h_2}-1} \right\} .$$

19 This indicates that, in practice, the hybrid algorithm assumes a scaled DGP  
 20  $p_\theta(\mathbf{x}|y)$  which scales the generative DGP  $p(\mathbf{x}|y)$  by a function not only of the  
 21 class label  $y$  but also of the feature vector  $\mathbf{x}$ .

1 *2.2 Class-posterior Probabilities*

2 The second perspective for interpreting the hybrid algorithm is via its mod-  
 3 elling of class-posterior probabilities,

$$p_{\theta}(y|\mathbf{x}) = \frac{p_{\theta}(\mathbf{x}, y)}{p_{\theta}(\mathbf{x})} = \frac{p(y)w_{\theta}(\mathbf{x}^1, \mathbf{x}^2)p(\mathbf{x}^1|y)^{\frac{\theta_1}{h_1}}p(\mathbf{x}^2|y)^{\frac{\theta_2}{h_2}}}{p_{\theta}(\mathbf{x})} = \frac{p(y)p(\mathbf{x}^1|y)^{\frac{\theta_1}{h_1}}p(\mathbf{x}^2|y)^{\frac{\theta_2}{h_2}}}{p_{\theta}(\mathbf{x})/w_{\theta}(\mathbf{x}^1, \mathbf{x}^2)},$$

4 where  $p_{\theta}(\mathbf{x}) = \sum_y p_{\theta}(\mathbf{x}, y) = p_{\theta}(\mathbf{x}, y = 1) + p_{\theta}(\mathbf{x}, y = 2)$ . According to Bayes'  
 5 Theorem, the class-posterior probabilities in terms of generative DGP  $p(\mathbf{x}|y)$   
 6 are  $p(y|\mathbf{x}) = p(y)p(\mathbf{x}|y)/p(\mathbf{x})$ ; it follows that

$$p_{\theta}(y|\mathbf{x}) = p(y|\mathbf{x}) \left\{ w_{\theta}(\mathbf{x}^1, \mathbf{x}^2) p(\mathbf{x}^1|y)^{\frac{\theta_1}{h_1}-1} p(\mathbf{x}^2|y)^{\frac{\theta_2}{h_2}-1} \frac{p(\mathbf{x})}{p_{\theta}(\mathbf{x})} \right\}.$$

7 This indicates that the normalised hybrid algorithm assumes scaled class-  
 8 posterior probabilities  $p_{\theta}(y|\mathbf{x})$  which scale the posterior probabilities  $p(y|\mathbf{x})$   
 9 by a function not only of the feature vector  $\mathbf{x}$  but also of the class label  $y$ .

10 *2.3 Loss Functions*

11 In order to find the best classifier, one of the optimal criteria is to minimize  
 12 the so-called unconditional or total risk:

$$R(\hat{y}) = E_y \left[ E_{\mathbf{x}|y} [L(y, \hat{y}(\mathbf{x}))] \right] = E_{\mathbf{x}} \left[ E_{y|\mathbf{x}} [L(y, \hat{y}(\mathbf{x}))] \right].$$

13 Such a criterion suffices to minimize the Bayes error, also called Bayes risk,

$$E_{y|\mathbf{x}} [L(y, \hat{y}(\mathbf{x}))] = \sum_{y=1}^K p(y|\mathbf{x}) L(y, \hat{y}(\mathbf{x})).$$

14 A simple and widely used loss function is a 0–1 loss such that  $L(y, \hat{y}(\mathbf{x})) = 1$  if  
 15  $\hat{y} \neq y$  and 0 otherwise. This leads to a Bayes classifier,  $\hat{y}(\mathbf{x}) = \operatorname{argmax}_y p(y|\mathbf{x})$ .

1 Since there are many loss functions that can lead to the normalised hybrid  
 2 algorithm, here we only present one loss function, fixing  $L(y, \hat{y}(\mathbf{x})) = 0$  if  
 3  $\hat{y} = y$ .

4 **Proposition 1** *If the number of groups is  $K \geq 2$ , and it is assumed that,*  
 5 *given  $y$ ,  $L(y, \hat{y}(\mathbf{x})) = L_y$  is independent of  $\hat{y}(\mathbf{x})$  if  $\hat{y} \neq y$ , then the hybrid*  
 6 *algorithm proposed in Raina et al. (2003) can be obtained through minimising*  
 7 *the Bayes error with a loss function  $L(y, \hat{y}(\mathbf{x}))$  such that  $L(y, \hat{y}(\mathbf{x})) = L_y$  if*  
 8  *$\hat{y} \neq y$  and 0 otherwise, where*

$$L_y = \frac{p(\mathbf{x}^1|y)^{\frac{\theta_1}{h_1}} p(\mathbf{x}^2|y)^{\frac{\theta_2}{h_2}}}{p(\mathbf{x}|y)},$$

9 *in which  $h_1$  and  $h_2$  are the dimensions of  $\mathbf{x}^1$  and  $\mathbf{x}^2$ , and  $\mathbf{x} = (\mathbf{x}^{1T}, \mathbf{x}^{2T})^T$ . A*  
 10 *generalisation of such a loss function is  $L_y = \frac{p_\theta(\mathbf{x}|y)}{p(\mathbf{x}|y)}$ .*

11 **PROOF.** The criterion to minimise the Bayes error for a classifier  $\hat{y}(\mathbf{x})$  with  
 12 such a loss function  $L(y, \hat{y}(\mathbf{x}))$  is

$$\begin{aligned} \hat{y}(\mathbf{x}) &= \operatorname{argmin}_{\hat{y}} \sum_{y \neq \hat{y}} p(y|\mathbf{x}) L_y = \operatorname{argmin}_{\hat{y}} \sum_{y \neq \hat{y}} p(\mathbf{x}^1|y)^{\frac{\theta_1}{h_1}} p(\mathbf{x}^2|y)^{\frac{\theta_2}{h_2}} p(y) \\ &= \operatorname{argmin}_{\hat{y}} -p(\mathbf{x}^1|\hat{y})^{\frac{\theta_1}{h_1}} p(\mathbf{x}^2|\hat{y})^{\frac{\theta_2}{h_2}} p(\hat{y}) = \operatorname{argmax}_y p(\mathbf{x}^1|y)^{\frac{\theta_1}{h_1}} p(\mathbf{x}^2|y)^{\frac{\theta_2}{h_2}} p(y), \end{aligned}$$

14 which is *Criterion-H*. The proof for the generalisation of  $L_y$  can be obtained  
 15 similarly by replacing  $p(\mathbf{x}^1|y)^{\frac{\theta_1}{h_1}} p(\mathbf{x}^2|y)^{\frac{\theta_2}{h_2}}$  with  $p_\theta(\mathbf{x}|y)$ .  $\square$

16 From Proposition 1, we observe that the loss from misclassification by the  
 17 hybrid algorithm depends on the accuracy of the approximation of the true  
 18 DGP  $p(\mathbf{x}|y)$  by the assumed one,  $p_\theta(\mathbf{x}|y)$  say. The closer  $p_\theta(\mathbf{x}|y)$  is to  $p(\mathbf{x}|y)$ ,  
 19 the closer  $L(y, \hat{y}(\mathbf{x}))$  can be approximated by a 0–1 loss function. Furthermore,  
 20 in contrast to the 0 – 1 loss,  $L_y$  is dependent on  $\mathbf{x}$ .

## 1 2.4 A Multi-class Extension

2 Interestingly, Fujino et al. (2007) present the result of a multi-class and multi-  
3 partition extension of the hybrid algorithm by maximising a conditional en-  
4 tropy of  $p(y|\mathbf{x})$  under some constraints associated with empirical joint distrib-  
5 ution  $p(\mathbf{x}, y)$  and class-conditional probabilities  $p(\mathbf{x}^r|y)$  for each partial feature  
6 vector  $\mathbf{x}^r, r = 1, \dots, R$ , as

$$p(y|\mathbf{x}) = \frac{e^{\mu_y} \prod_{r=1}^R p(\mathbf{x}^r|y)^{\lambda_r}}{\sum_y e^{\mu_y} \prod_{r=1}^R p(\mathbf{x}^r|y)^{\lambda_r}},$$

7 where  $\lambda_r$  and  $\mu_y$  are the Lagrange multipliers. This result is equivalent to a  
8 straightforward extension of the hybrid algorithm, in which  $\lambda_r = \theta_r/h_r$  and  
9  $\mu_y = \log p(y) + \log w_\theta(\mathbf{x})$ .

## 10 3 Simulation Studies

### 11 3.1 Parameter Estimation, Implementation and Evaluation of the Classi- 12 fiers, and Simulated Datasets

#### 13 3.1.1 Discriminative Learning of $\theta$

14 By “hybrid”, the normalised hybrid algorithm proposed in Raina et al. (2003)  
15 means to use a discriminative approach to the estimation of  $\theta$  such that

$$\hat{\theta} = \operatorname{argmax}_{\theta} \sum_{i=1}^m \log p_{\theta}(y^{(i)}|\mathbf{x}^{(i)}) = \operatorname{argmax}_{\theta} \sum_{i=1}^m \log \frac{p_{\theta}(\mathbf{x}^{(i)}, y^{(i)})}{\sum_y p_{\theta}(\mathbf{x}^{(i)}, y)},$$

16 where  $m$  is the number of independent training samples  $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^m$ , in  
17 which  $(\mathbf{x}^{(i)})^T = ((\mathbf{x}^{1,(i)})^T, (\mathbf{x}^{2,(i)})^T)$ . If  $y$  is a binary variable such that  $y \in$

1  $\{1, 2\}$ ,  $p_\theta(y = 1|\mathbf{x})$  can be written in a way similar to that of logistic regression:

$$p_\theta(y = 1|\mathbf{x}) = \frac{\exp(\lambda_H(\mathbf{x}))}{1 + \exp(\lambda_H(\mathbf{x}))} ,$$

2 where  $\lambda_H(\mathbf{x})$ , as defined in Section 2.1, is the discriminant function corre-  
 3 sponding to *Criterion-H*. As with linear logistic regression,  $\lambda_H(\mathbf{x})$  is a linear  
 4 function of  $\theta_1$  and  $\theta_2$ .

5 Instead of using maximisation, we minimise negative loglikelihood  $-\ell_H$  to  
 6 estimate  $\theta_1$  and  $\theta_2$ , where

$$-\ell_H = -\sum_{i=1}^m \log p_\theta(y^{(i)}|\mathbf{x}^{(i)})$$

7

$$= \sum_{i=1}^m \left\{ \mathbf{1}_{\{y^{(i)}=1\}} \log \left( 1 + e^{-\lambda_H(\mathbf{x}^{(i)})} \right) + \mathbf{1}_{\{y^{(i)}=2\}} \log \left( 1 + e^{\lambda_H(\mathbf{x}^{(i)})} \right) \right\} .$$

8 Concerning  $\lambda_H(\mathbf{x})$ , in order to estimate the parameters in the same discrim-  
 9 inative way as that of linear logistic regression, Raina et al. (2003) redefine  
 10  $\theta$  as  $\theta = (\theta_0, \theta_1, \theta_2)^T$ , where  $\theta_0 = \log \frac{p(y=1)}{p(y=2)}$ , as the intercept in a linear lo-  
 11 gistic regression model, is estimated discriminatively, *i.e.*,  $\log \frac{p(y=1)}{p(y=2)}$  is not  
 12 calculated by using generative estimators of  $p(y = 1)$  and  $p(y = 2)$  but is  
 13 directly estimated by a discriminative approach. Except for that,  $\log \frac{p(\mathbf{x}^1|y=1)}{p(\mathbf{x}^1|y=2)}$   
 14 and  $\log \frac{p(\mathbf{x}^2|y=1)}{p(\mathbf{x}^2|y=2)}$  are estimated by a generative approach.

15 Considering that the discriminative estimator of  $\theta$  uses outputs of the genera-  
 16 tive estimator of  $p(\mathbf{x}|y)$  as inputs while both estimators use the same training  
 17 set, Raina et al. (2003) suggest that the discriminative estimator of  $\theta$  is biased.

18 Consequently, they use a “leave-one-out” strategy as follows:

$$\hat{\theta}_{-i} = \operatorname{argmax}_{\theta} \sum_{i=1}^m \log p_{\theta,-i}(y^{(i)}|\mathbf{x}^{(i)}) = \operatorname{argmax}_{\theta} \sum_{i=1}^m \log \frac{p_{\theta,-i}(\mathbf{x}^{(i)}, y^{(i)})}{\sum_y p_{\theta,-i}(\mathbf{x}^{(i)}, y)} ,$$

19 where  $p_{\theta,-i}(\mathbf{x}^{(i)}, y)$  and  $p_{\theta,-i}(\mathbf{x}^{(i)}, y^{(i)})$  are obtained from the data with the  $i$ -th

1 observation removed. However, when the training set size  $m$  is large enough,  
 2 there is little difference between  $\hat{\theta}_{-i}$  and  $\hat{\theta}$ , and thus such a bias can be ignored.  
 3 Therefore, in our study, we do not use the “leave-one-out” strategy to estimate  
 4  $\theta$ .

### 5 3.1.2 Implementation of the Classifiers

6 In order to evaluate the discrimination performance of the hybrid algorithm,  
 7 we compare it with two widely-used discriminative and generative classifiers,  
 8 linear logistic regression and the naïve Bayes classifier, using simulated con-  
 9 tinuous and discrete data.

10 The naïve Bayes classifier is implemented by an R function *naiveBayes* from  
 11 a contributed package **e1071** for R. As with Raina et al. (2003), for discrete  
 12 data, we use Laplace (add-one) smoothing. For simulated continuous data, the  
 13 naïve Bayes classifier, which assumes normal distributions for class-conditional  
 14 probabilities  $p(\mathbf{x}|y)$ , corresponds to LDA- $\Lambda$  when the covariance matrix  $\Sigma_1$  of  
 15 the group  $y = 1$  is equal to the covariance matrix  $\Sigma_2$  of the group  $y = 2$ ,  
 16 and corresponds to quadratic normal discriminant analysis with a common  
 17 diagonal covariance matrix (QDA- $\Lambda$ ) when  $\Sigma_1 \neq \Sigma_2$ . The naïve Bayes classifier  
 18 assumes the conditional independence of all  $h$  features given the group label  
 19  $y$ , such that  $p(\mathbf{x}|y) = \prod_{j=1}^h p(x_j|y)$ ; its discriminant function  $\lambda_G(\mathbf{x})$  can be  
 20 written as

$$\lambda_G(\mathbf{x}) = \log \frac{p(y=1)}{p(y=2)} + \sum_{j=1}^h \log \frac{p(x_j|y=1)}{p(x_j|y=2)} .$$

21 The implementation of parameter estimation for the hybrid algorithm with  
 22  $\lambda_H(\mathbf{x})$  consists of two steps: in first step, by use of the R function *naive-*  
 23 *Bayes*,  $p(x_j|y)$ ,  $j = 1, \dots, h$ , are generatively estimated and thus  $\log \frac{p(\mathbf{x}^1|y=1)}{p(\mathbf{x}^1|y=2)}$

1 and  $\log \frac{p(\mathbf{x}^2|y=1)}{p(\mathbf{x}^2|y=2)}$  can be calculated; in second step,  $\theta$  is estimated discrimina-  
 2 tively by use of an R function *glm* (from a standard package **stats** in R) with  
 3  $\log \frac{p(\mathbf{x}^1|y=1)}{p(\mathbf{x}^1|y=2)}$  and  $\log \frac{p(\mathbf{x}^2|y=1)}{p(\mathbf{x}^2|y=2)}$  as predictor variables. The hybrid algorithm as-  
 4 sumes conditional independence within the partial feature vectors such that  
 5  $p(\mathbf{x}^1|y) = \prod_{j=1}^{h_1} p(x_j|y)$  and  $p(\mathbf{x}^2|y) = \prod_{j=h_1+1}^h p(x_j|y)$ .

6 Linear logistic regression is implemented by the R function *glm* which uses  
 7 an iteratively reweighted least squares algorithm (IRLS, or IWLS, also known  
 8 as the Fisher scoring algorithm) to fit the model. The discriminant function  
 9  $\lambda_D(\mathbf{x})$  of linear logistic regression can be written as

$$\lambda_D(\mathbf{x}) = \beta_0 + \sum_{j=1}^h \beta_j x_j ,$$

10 which does not necessarily imply that the conditional independence assump-  
 11 tion holds.

### 12 3.1.3 Evaluation of the Classifiers

13 To evaluate the performance of the three classifiers, we use the misclassification  
 14 error rate (ER) and logistic loss (LL). The ER is defined as usual by the  
 15 number of misclassified observations over the total number of observations; it  
 16 is based on a 0 – 1 loss function and is independent of the observed value  $x$ .

17 In contrast, the LL, resembling the soft margin loss used in machine learning,  
 18 is dependent on  $x$ . It is based on a loss function  $L(y, \hat{y}(\mathbf{x})) = -\log p(y|\mathbf{x})$  and  
 19 thus defined by

$$LL = \sum_{i=1}^m \left\{ -\log p(y^{(i)}|\mathbf{x}^{(i)}) \right\} .$$

20 It can be easily recognised that the LL is in fact the negative of the log-  
 21 likelihood of  $p(y|\mathbf{x})$ , and therefore the estimates obtained by the discrimina-

1 tive classifiers provide the best classification for the training samples if the  
 2 minimum LL is used to measure the performance.

3 Consider two groups  $y \in \{1, 2\}$  with the discriminant function  $\lambda(\mathbf{x}) = \log \frac{p(y=1|\mathbf{x})}{p(y=2|\mathbf{x})}$ .

4 Then the LL can be rewritten as

$$LL = \sum_{i=1}^m \left\{ \left( -\log \frac{e^{\lambda(\mathbf{x}^{(i)})}}{1 + e^{\lambda(\mathbf{x}^{(i)})}} \right)^{\mathbf{1}_{\{y^{(i)}=1\}}} \left( -\log \frac{1}{1 + e^{\lambda(\mathbf{x}^{(i)})}} \right)^{\mathbf{1}_{\{y^{(i)}=2\}}} \right\},$$

5 where  $\mathbf{1}_{\{y^{(i)}=k\}}$  is an indicator function of the subset  $\{y^{(i)} = k\}$ . A simple  
 6 notation for the LL used by the machine learning community for two groups  
 7 such that  $y \in \{-1, 1\}$  is

$$LL = \sum_{i=1}^m \left\{ -\log \frac{1}{1 + e^{-y^{(i)}\lambda(\mathbf{x}^{(i)})}} \right\} = \sum_{i=1}^n \left\{ \log \left( 1 + e^{-y^{(i)}\lambda(\mathbf{x}^{(i)})} \right) \right\}.$$

### 8 3.1.4 Simulated Datasets

9 In our studies, 12 datasets are simulated, of which 6 are composed of  $h$  con-  
 10 tinuous features and the other 6 are composed of  $h$  discrete features. In each  
 11 continuous dataset, the data arise from two  $h$ -variate normal distributions; in  
 12 each discrete dataset, the data arise from two  $h$ -variate Bernoulli distributions.

13 Each dataset consists of  $10^3$  observations, which are equally categorised into  
 14 two groups by a group label  $y \in 1, 2$ . Amongst them,  $m/2$  observations from  
 15 each of the two groups are used as training samples;  $m$  is sampled within  
 16  $[100, 400]$  in steps of 25. For each sampled  $m$ , the  $10^3$  observations are ran-  
 17 domly split into  $m$  training samples and  $10^3 - m$  test samples with 400 repli-  
 18 cates; from them, the medians of the ERs and LLs are recorded and plotted.

19 In each dataset, we set  $h = 4$  and the feature vector  $\mathbf{x} = (x_1, x_2, x_3, x_4)^T$  is  
 20 composed of 2 partial feature vectors  $\mathbf{x}^1 = (x_1, x_2)^T$  and  $\mathbf{x}^2 = (x_3, x_4)^T$ , *i.e.*,

1  $h_1 = h_2 = 2$ .

2 Amongst the 12 datasets, 6 datasets (3 continuous and 3 discrete) have  $\Sigma_1 =$   
3  $\Sigma_2$ , *i.e.*, the two groups have a common covariance matrix  $\Sigma$ . In addition,  
4 there are 4 datasets (2 continuous and 2 discrete) having diagonal covariance  
5 matrices, and thus for them the assumption of conditional independence of all  
6  $h$  features of  $\mathbf{x}$  given  $y$  underlying the naïve Bayes classifier is satisfied. There  
7 are also 4 datasets having block diagonal covariance matrices of two blocks,  
8 where one block consists of the  $h_1$  features of  $\mathbf{x}^1$  and the other one consists  
9 of the  $h_2$  features of  $\mathbf{x}^2$ , and thus for them the assumption of conditional  
10 independence between  $\mathbf{x}^1$  and  $\mathbf{x}^2$  given  $y$  is satisfied. The other 4 datasets  
11 have full covariance matrices such that each of the  $h$  features of  $\mathbf{x}$  given  $y$  is  
12 dependent on the others.

13 *3.2 Results for Simulated Continuous Data with a Common Covariance Ma-*  
14 *trix  $\Sigma$*

15 The first 3 datasets contain simulated continuous data arising from two 4-  
16 variate normal distributions:  $\mathbf{x} \sim \mathcal{N}(\mu_1, \Sigma_1)$  for the group with  $y = 1$  and  
17  $\mathbf{x} \sim \mathcal{N}(\mu_2, \Sigma_2)$  for  $y = 2$  with  $\mu_1 = (1.5, 0, 0.5, 0)^T$ ,  $\mu_2 = (-1.5, 0, -0.5, 0)^T$ ,  
18  $\Sigma_1 = \Sigma_2 = \Sigma$  and  $\Sigma$  is

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} 1 & c & 0 & 0 \\ c & 1 & 0 & 0 \\ 0 & 0 & 1 & c \\ 0 & 0 & c & 1 \end{bmatrix} \text{ or } \begin{bmatrix} 1 & c & c & c \\ c & 1 & c & c \\ c & c & 1 & c \\ c & c & c & 1 \end{bmatrix}$$

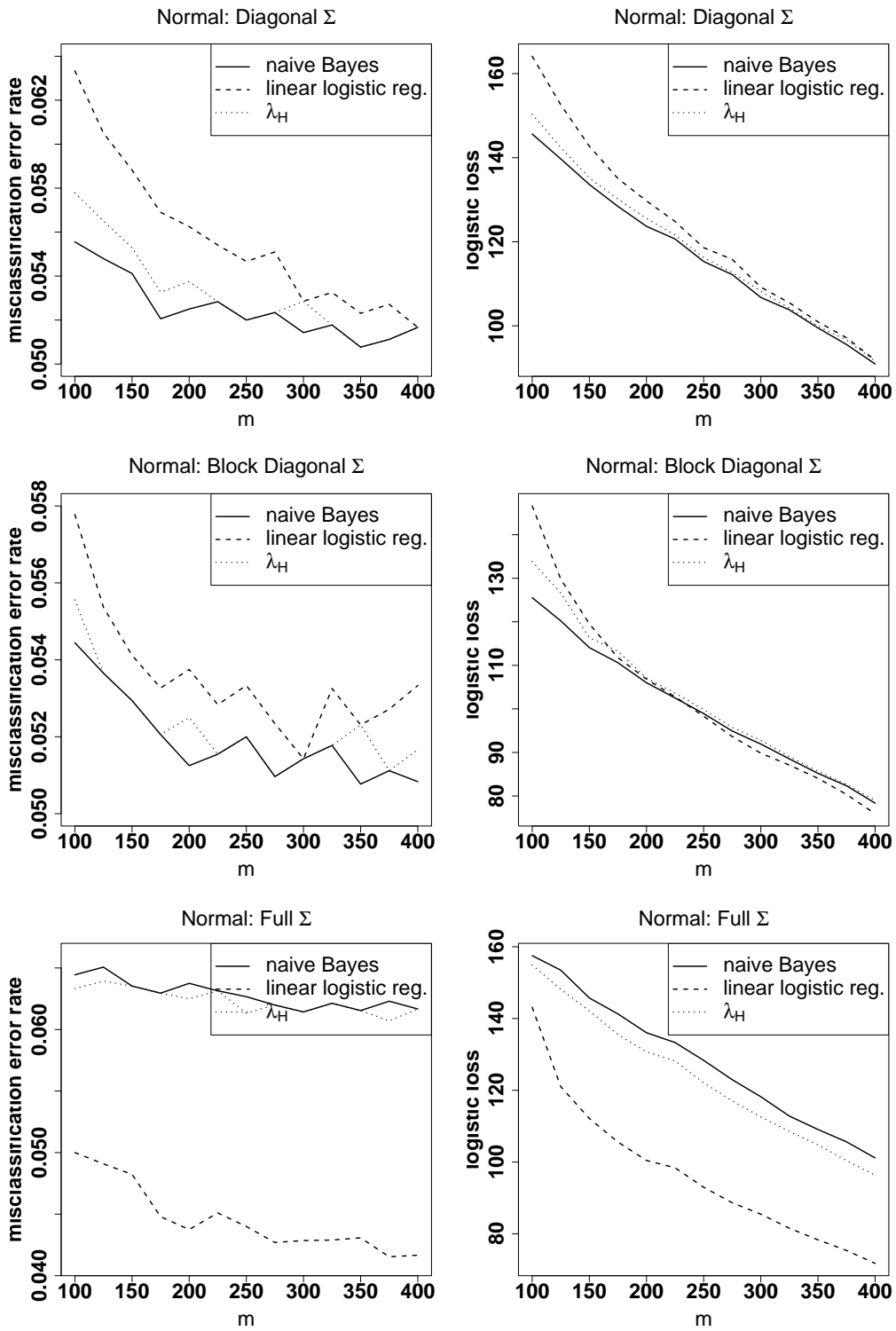


Figure 1. Simulated normally distributed data with equal covariance matrices. Plots of classification performance measured by ER and by LL vs. training set size  $m$ .

1 with  $c = 0.25$ , giving a diagonal, a block diagonal and a full covariance matrix,  
 2 respectively, for the 3 datasets.

3 Medians of the ERs and LLs are obtained from 400 replicates; the medians  
 4 are plotted against the training set size  $m$  in Figure 1, of which each row  
 5 represents the results for one dataset.

### 6 3.3 Results for Simulated Continuous Data with Unequal Covariance Matri-

7 ces  $\Sigma_1, \Sigma_2$

8 The setting of the second set of 3 datasets is similar to that of the first set in  
 9 Section 3.2, except that  $\Sigma_1 \neq \Sigma_2$  and  $\Sigma_2$  is

$$\begin{bmatrix} 0.25 & 0 & 0 & 0 \\ 0 & 0.75 & 0 & 0 \\ 0 & 0 & 1.25 & 0 \\ 0 & 0 & 0 & 1.75 \end{bmatrix}, \begin{bmatrix} 0.25 & c & 0 & 0 \\ c & 0.75 & 0 & 0 \\ 0 & 0 & 1.25 & c \\ 0 & 0 & c & 1.75 \end{bmatrix} \text{ or } \begin{bmatrix} 0.25 & c & c & c \\ c & 0.75 & c & c \\ c & c & 1.25 & c \\ c & c & c & 1.75 \end{bmatrix}$$

10 while  $\Sigma_1$  is the same as  $\Sigma$  shown in Section 3.2, respectively for these 3  
 11 datasets.

12 The results for these 3 datasets are shown in Figure 2.

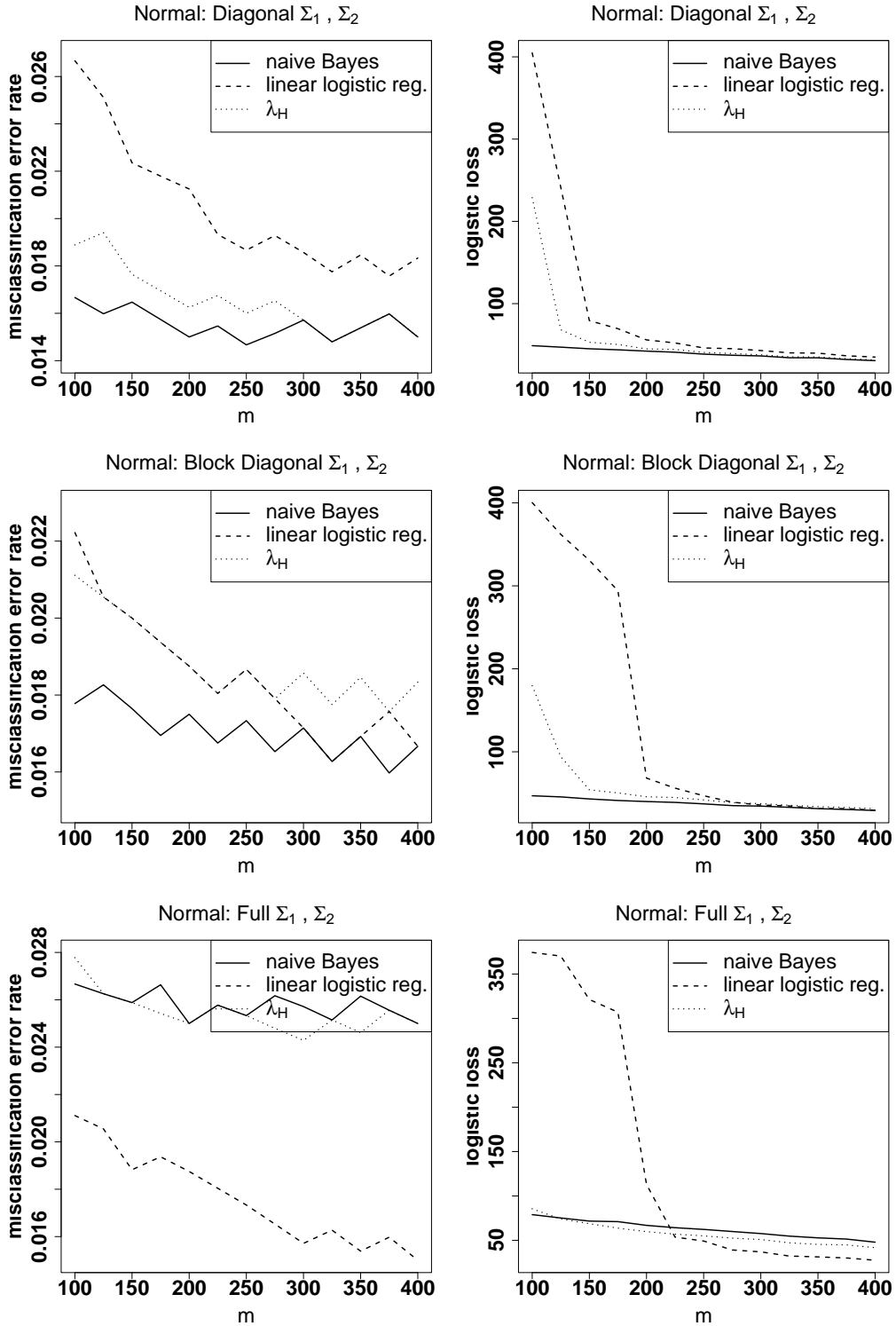


Figure 2. Simulated normally distributed data with unequal covariance matrices. Plots of classification performance measured by ER and by LL vs. training set size  $m$ .

1 3.4 Results for Simulated Discrete Data with a Common Covariance Matrix

2  $\Sigma$

3 The third set of 3 datasets contains simulated discrete data arising from two  
 4 4-variate Bernoulli distributions:  $\mathbf{x} \sim B(\mathbf{p})$  for the group with  $y = 1$  and  
 5  $\mathbf{x} \sim B(\mathbf{q})$  for  $y = 2$ , where  $\mathbf{p} = (p_1, p_2, p_3, p_4)^T = (0.2, 0.3, 0.4, 0.5)^T$ ,  $\mathbf{q} =$   
 6  $(q_1, q_2, q_3, q_4)^T = (0.8, 0.7, 0.6, 0.5)^T$ . In this context, the two groups have a  
 7 common covariance matrix  $\Sigma$  but different means ( $\mu_1 = E\{\mathbf{x}|y = 1\} = \mathbf{p}$  and  
 8  $\mu_2 = E\{\mathbf{x}|y = 2\} = \mathbf{q}$ ).  $\Sigma$  is a diagonal, block diagonal and full covariance  
 9 matrix, respectively for these 3 datasets.

10 For the first dataset, each of the 4 features  $\{x_j\}_{j=1}^4$  is conditionally independent  
 11 of the others given the group label  $y$ . In order to achieve this, we set all the  
 12 elements of  $\mathbf{p}$  and  $\mathbf{q}$  such that the covariance matrices for the two groups are  
 13 diagonal matrices:

$$\Sigma_{y=1} = \text{diag}(V_{1,1}, V_{2,2}, V_{3,3}, V_{4,4}) , \quad \Sigma_{y=2} = \text{diag}(V'_{1,1}, V'_{2,2}, V'_{3,3}, V'_{4,4}) ,$$

14 where, for  $i = 1, \dots, 4$ ,

$$V_{i,i} = p_i(1 - p_i) , \quad V'_{i,i} = q_i(1 - q_i) .$$

15 In order to have  $\Sigma_{y=1} = \Sigma_{y=2} = \Sigma$ , we set  $q_i = 1 - p_i$ .

16 For the second dataset,  $\mathbf{x}^1$  is conditionally independent of  $\mathbf{x}^2$  given the group  
 17 label  $y$ . In order to achieve this, we set only  $p_1, p_3, q_1, q_3$  and conditional prob-  
 18 abilities  $p_{2|1(1)}, p_{2|1(0)}, p_{4|3(1)}, p_{4|3(0)}$  and  $q_{2|1(1)}, q_{2|1(0)}, q_{4|3(1)}, q_{4|3(0)}$ , where  $p_{i|j(v)}$   
 19 and  $q_{i|j(v)}$  denote the success probabilities  $p_i$  and  $q_i$  of  $x_i$  given  $x_j = v, v \in 0, 1$ .

1 It follows that

$$p_2 = p_1 p_{2|1(1)} + (1 - p_1) p_{2|1(0)} , \quad q_2 = q_1 q_{2|1(1)} + (1 - q_1) q_{2|1(0)} ,$$

2

$$p_4 = p_3 p_{4|3(1)} + (1 - p_3) p_{4|3(0)} , \quad q_4 = q_3 q_{4|3(1)} + (1 - q_3) q_{4|3(0)} ,$$

3 and the covariance matrices for the two groups are block diagonal, symmetric  
4 matrices:

$$\Sigma_{y=1} = \begin{bmatrix} V_{1,1} & V_{1,2} & 0 & 0 \\ V_{1,2} & V_{2,2} & 0 & 0 \\ 0 & 0 & V_{3,3} & V_{3,4} \\ 0 & 0 & V_{3,4} & V_{4,4} \end{bmatrix} , \quad \Sigma_{y=2} = \begin{bmatrix} V'_{1,1} & V'_{1,2} & 0 & 0 \\ V'_{1,2} & V'_{2,2} & 0 & 0 \\ 0 & 0 & V'_{3,3} & V'_{3,4} \\ 0 & 0 & V'_{3,4} & V'_{4,4} \end{bmatrix} ,$$

5 where, for  $i = 1, \dots, 4$ ,

$$V_{i,i} = p_i (1 - p_i) , \quad V'_{i,i} = q_i (1 - q_i) ,$$

6

$$V_{1,2} = p_1 (p_{2|1(1)} - p_2) , \quad V'_{1,2} = q_1 (q_{2|1(1)} - q_2) ,$$

7

$$V_{3,4} = p_3 (p_{4|3(1)} - p_4) , \quad V'_{3,4} = q_3 (q_{4|3(1)} - q_4) .$$

8 In order to have  $\Sigma_{y=1} = \Sigma_{y=2} = \Sigma$ , we set

$$q_1 = 1 - p_1 , \quad q_3 = 1 - p_3 ,$$

9

$$q_{2|1(1)} = 1 - p_{2|1(0)} , \quad q_{2|1(0)} = 1 - p_{2|1(1)} , \quad \text{and}$$

10

$$q_{4|3(1)} = 1 - p_{4|3(0)} , \quad q_{4|3(0)} = 1 - p_{4|3(1)} .$$

11 For the third dataset, each of the 4 features  $\{x_j\}_{j=1}^4$  is dependent on the  
12 others given the group label  $y$ . In order to achieve that, we set only  $p_1, q_1$  and  
13 conditional probabilities  $p_{i|1(1)}, p_{i|1(0)}$  and  $q_{i|1(1)}, q_{i|1(0)}$ , for  $i = 2, 3, 4$ . It follows

1 that, for  $i = 2, 3, 4$ ,

$$p_i = p_1 p_{i|1(1)} + (1 - p_1) p_{i|1(0)} , \quad q_i = q_1 q_{i|1(1)} + (1 - q_1) q_{i|1(0)} ,$$

2 and the covariance matrices for the two groups are full symmetric matrices:

$$\Sigma_{y=1} = \begin{bmatrix} V_{1,1} & V_{1,2} & V_{1,3} & V_{1,4} \\ V_{1,2} & V_{2,2} & V_{2,3} & V_{2,4} \\ V_{1,3} & V_{2,3} & V_{3,3} & V_{3,4} \\ V_{1,4} & V_{2,4} & V_{3,4} & V_{4,4} \end{bmatrix} , \quad \Sigma_{y=2} = \begin{bmatrix} V'_{1,1} & V'_{1,2} & V'_{1,3} & V'_{1,4} \\ V'_{1,2} & V'_{2,2} & V'_{2,3} & V'_{2,4} \\ V'_{1,3} & V'_{2,3} & V'_{3,3} & V'_{3,4} \\ V'_{1,4} & V'_{2,4} & V'_{3,4} & V'_{4,4} \end{bmatrix} ,$$

3 where

$$V_{i,i} = p_i (1 - p_i) , \quad V'_{i,i} = q_i (1 - q_i) , \quad i = 1, \dots, 4 ;$$

$$V_{1,i} = p_1 (p_{i|1(1)} - p_i) , \quad V'_{1,i} = q_1 (q_{i|1(1)} - q_i) , \quad i = 2, 3, 4 ;$$

5 and, for  $i, j = 2, 3, 4$ ,

$$p(x_i = 1, x_j = 1) = p_1 p_{i|1(1)} p_{j|1(1)} + (1 - p_1) p_{i|1(0)} p_{j|1(0)} ,$$

$$q(x_i = 1, x_j = 1) = q_1 q_{i|1(1)} q_{j|1(1)} + (1 - q_1) q_{i|1(0)} q_{j|1(0)} ,$$

7 such that

$$V_{i,j} = p(x_i = 1, x_j = 1) - p_i p_j , \quad V'_{i,j} = q(x_i = 1, x_j = 1) - q_i q_j .$$

8 In order to have  $\Sigma_{y=1} = \Sigma_{y=2} = \Sigma$ , we set

$$q_1 = 1 - p_1 ,$$

$$q_{i|1(1)} = 1 - p_{i|1(0)} , \quad \text{and} \quad q_{i|1(0)} = 1 - p_{i|1(1)} , \quad i = 2, 3, 4 .$$

1 3.4.1 Diagonal Covariance Matrix  $\Sigma$

2 For the first dataset, we set  $\mu_1 = \mathbf{p} = (0.2, 0.3, 0.4, 0.5)^T$ ,  $\mu_2 = \mathbf{q} = \mathbf{1} - \mathbf{p} =$   
3  $(0.8, 0.7, 0.6, 0.5)^T$  such that the common covariance matrix  $\Sigma$  is a diagonal  
4 matrix,  $\text{diag}(0.16, 0.21, 0.24, 0.25)$ .

5 3.4.2 Block Diagonal Covariance Matrix  $\Sigma$

6 For the second dataset, we set

$$p_1 = 0.2, q_1 = 1 - p_1 = 0.8,$$

7

$$p_3 = 0.4, q_3 = 1 - p_3 = 0.6;$$

8

$$p_{2|1(1)} = 0.7, p_{2|1(0)} = 0.2,$$

9

$$q_{2|1(1)} = 1 - p_{2|1(0)} = 0.8, q_{2|1(0)} = 1 - p_{2|1(1)} = 0.3;$$

10

$$p_{4|3(1)} = 0.8, p_{4|3(0)} = 0.3,$$

11

$$q_{4|3(1)} = 1 - p_{4|3(0)} = 0.7, \text{ and } q_{4|3(0)} = 1 - p_{4|3(1)} = 0.2.$$

12 It follows that  $\mu_1 = \mathbf{p} = (0.2, 0.3, 0.4, 0.5)^T$ ,  $\mu_2 = \mathbf{q} = \mathbf{1} - \mathbf{p} = (0.8, 0.7, 0.6, 0.5)^T$ ,

13 and  $\Sigma$  is a block diagonal matrix 
$$\begin{bmatrix} 0.16 & 0.08 & 0 & 0 \\ 0.08 & 0.21 & 0 & 0 \\ 0 & 0 & 0.24 & 0.12 \\ 0 & 0 & 0.12 & 0.25 \end{bmatrix}.$$

1 *3.4.3 Full Covariance Matrix  $\Sigma$*

2 For the third dataset, we set

$$p_1 = 0.2, q_1 = 1 - p_1 = 0.8;$$

3

$$p_{2|1(1)} = 0.7, p_{2|1(0)} = 0.2,$$

4

$$q_{2|1(1)} = 1 - p_{2|1(0)} = 0.8, q_{2|1(0)} = 1 - p_{2|1(1)} = 0.3;$$

5

$$p_{3|1(1)} = 0.8, p_{3|1(0)} = 0.3,$$

6

$$q_{3|1(1)} = 1 - p_{3|1(0)} = 0.7, q_{3|1(0)} = 1 - p_{3|1(1)} = 0.2;$$

7

$$p_{4|1(1)} = 0.9, p_{4|1(0)} = 0.4,$$

8

$$q_{4|1(1)} = 1 - p_{4|1(0)} = 0.6, \text{ and } q_{4|1(0)} = 1 - p_{4|1(1)} = 0.1.$$

9 It follows that  $\mu_1 = \mathbf{p} = (0.2, 0.3, 0.4, 0.5)^T$ ,  $\mu_2 = \mathbf{q} = \mathbf{1} - \mathbf{p} = (0.8, 0.7, 0.6, 0.5)^T$ ,

10 and  $\Sigma$  is a full matrix

$$\begin{bmatrix} 0.16 & 0.08 & 0.08 & 0.08 \\ 0.08 & 0.21 & 0.04 & 0.04 \\ 0.08 & 0.04 & 0.24 & 0.04 \\ 0.08 & 0.04 & 0.04 & 0.25 \end{bmatrix}.$$

11 The results for these 3 datasets are shown in Figure 3.

12 *3.5 Results for Simulated Discrete Data with Unequal Covariance Matrices*

13  $\Sigma_1, \Sigma_2$

14 The settings of the last 3 datasets are similar to those of the third set in

15 Section 3.4, except that  $\Sigma_1 \neq \Sigma_2$  and  $\mathbf{q}$  is different amongst these 3 datasets.

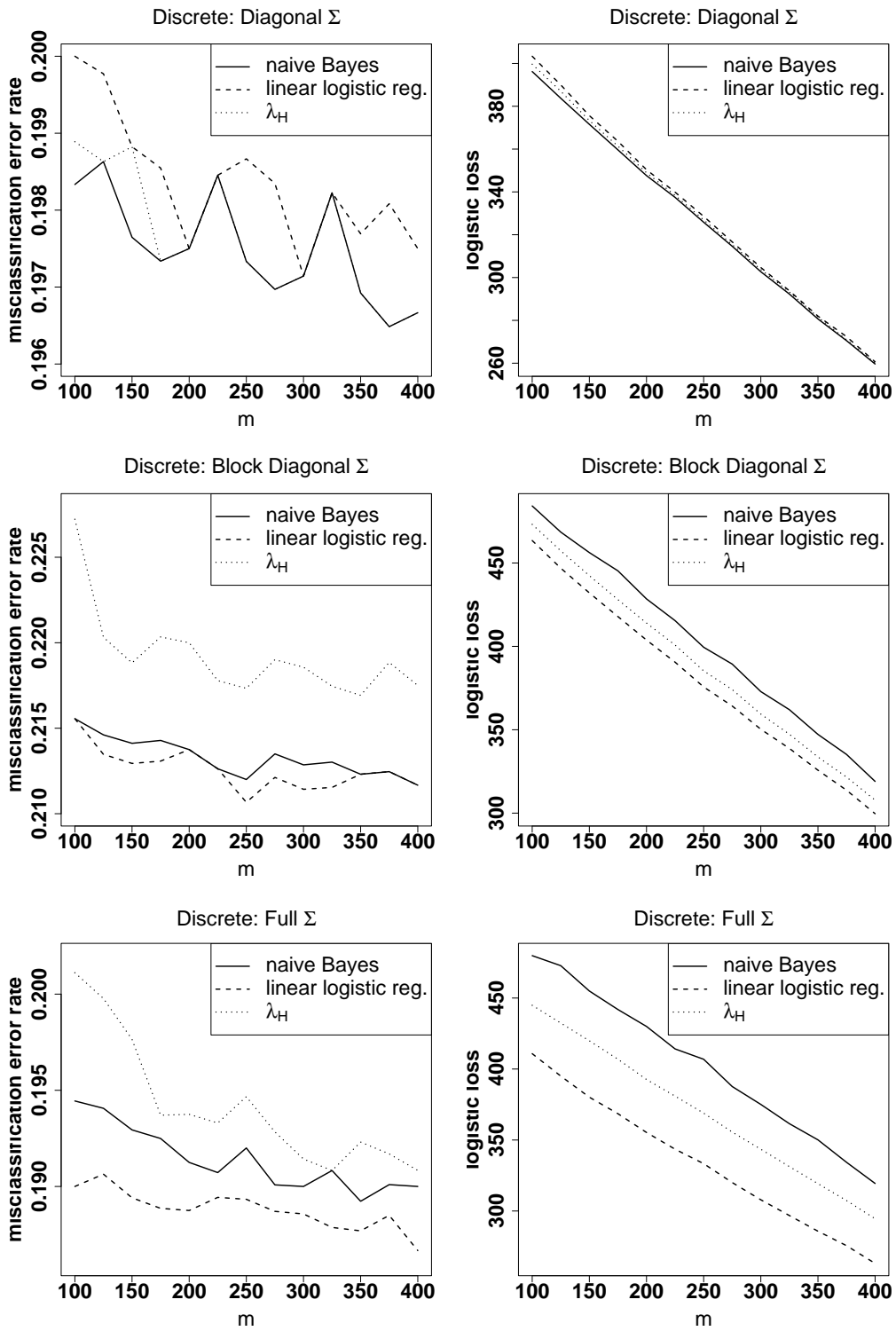


Figure 3. Simulated Bernoulli data with equal covariance matrices. Plots of classification performance measured by ER and by LL vs. training set size  $m$ .

1 3.5.1 Diagonal Covariance Matrices  $\Sigma_1, \Sigma_2$

2 For the first dataset, the setting is the same as that in Section 3.4.1 except that  
3  $\mathbf{q} = \mathbf{p} + 0.4$  rather than  $\mathbf{q} = \mathbf{1} - \mathbf{p}$ . That is, we set  $\mu_1 = \mathbf{p} = (0.2, 0.3, 0.4, 0.5)^T$ ,  
4  $\mu_2 = \mathbf{q} = (0.6, 0.7, 0.8, 0.9)^T$  such that  $\Sigma_1 = \text{diag}(0.16, 0.21, 0.24, 0.25)$  and  
5  $\Sigma_2 = \text{diag}(0.24, 0.21, 0.16, 0.09)$ .

6 3.5.2 Block Diagonal Covariance Matrices  $\Sigma_1, \Sigma_2$

7 For the second dataset, the setting is the same as that in Section 3.4.2 except  
8 that  $q_1 = p_1 + 0.4, q_3 = p_3 + 0.4$  rather than  $q_1 = 1 - p_1, q_3 = 1 - p_3$ , respectively.  
9 That is, we have  $\mu_1 = \mathbf{p} = (0.2, 0.3, 0.4, 0.5)^T, \mu_2 = \mathbf{q} = (0.6, 0.6, 0.8, 0.6)^T$ ,

10  $\Sigma_1 = \begin{bmatrix} 0.16 & 0.08 & 0 & 0 \\ 0.08 & 0.21 & 0 & 0 \\ 0 & 0 & 0.24 & 0.12 \\ 0 & 0 & 0.12 & 0.25 \end{bmatrix}$  and  $\Sigma_2 = \begin{bmatrix} 0.24 & 0.12 & 0 & 0 \\ 0.12 & 0.24 & 0 & 0 \\ 0 & 0 & 0.16 & 0.08 \\ 0 & 0 & 0.08 & 0.24 \end{bmatrix}$ .

11 3.5.3 Full Covariance Matrices  $\Sigma_1, \Sigma_2$

12 For the third dataset, the setting is the same as that in Section 3.4.3 except  
13 that  $q_1 = p_1 + 0.4$  rather than  $q_1 = 1 - p_1$ . That is, we have  $\mu_1 = \mathbf{p} =$

$$\begin{aligned}
& \mu_1 = (0.2, 0.3, 0.4, 0.5)^T, \mu_2 = \mathbf{q} = (0.6, 0.6, 0.5, 0.4)^T, \Sigma_1 = \begin{bmatrix} 0.16 & 0.08 & 0.08 & 0.08 \\ 0.08 & 0.21 & 0.04 & 0.04 \\ 0.08 & 0.04 & 0.24 & 0.04 \\ 0.08 & 0.04 & 0.04 & 0.25 \end{bmatrix} \\
& \text{and } \Sigma_2 = \begin{bmatrix} 0.24 & 0.12 & 0.12 & 0.12 \\ 0.12 & 0.24 & 0.06 & 0.06 \\ 0.12 & 0.06 & 0.25 & 0.06 \\ 0.12 & 0.06 & 0.06 & 0.24 \end{bmatrix}; \text{ they are symmetric, positive-definite matrices.}
\end{aligned}$$

The results for these 3 datasets are shown in Figure 4.

### 3.6 Summary of Simulation Studies

With the results shown in Figure 1, 2, 3 and 4, our simulation studies on both the continuous and discrete datasets suggest the follow conclusions.

- In general, in terms of both performance measures, namely ER and LL, if both the covariance matrices  $\Sigma_1$  and  $\Sigma_2$  are diagonal matrices, the naïve Bayes classifier performs the best; if both the covariance matrices  $\Sigma_1$  and  $\Sigma_2$  are non-diagonal matrices, linear logistic regression performs the best, in particular when the training set size  $m$  is large.
- With these simulated datasets, our studies cannot support the claim that the hybrid algorithm offers better performance than either the naïve Bayes

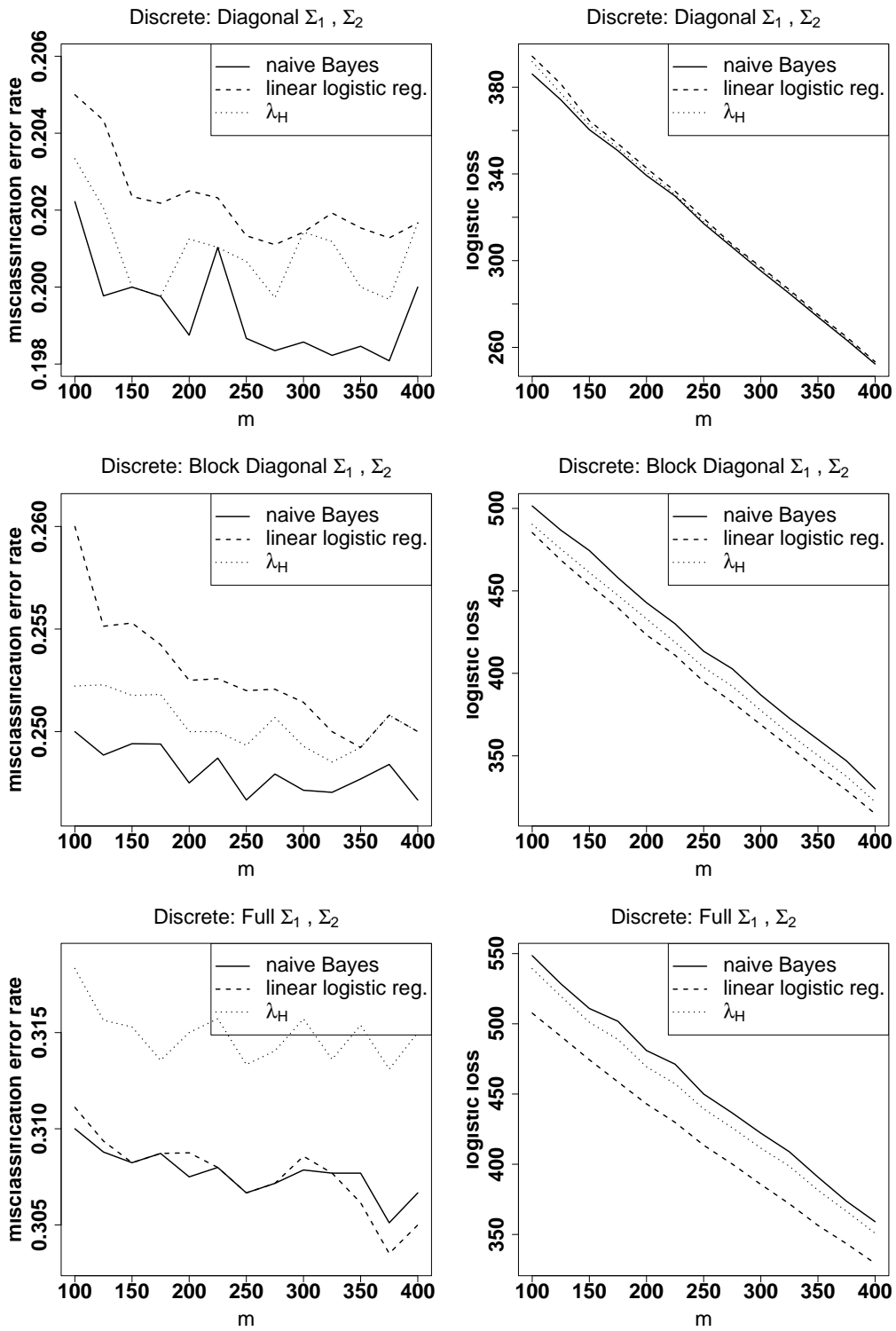


Figure 4. Simulated Bernoulli data with unequal covariance matrices. Plots of classification performance measured by ER and by LL vs. training set size  $m$ .

1 classifier or linear logistic regression alone, a phenomenon observed from  
2 the empirical studies in Raina et al. (2003).

3 The superior performance of the naïve Bayes classifier can be attributed to  
4 the fact that the simulated data satisfy the assumption of conditional inde-  
5 pendence underlying the classifier; the superior performance of linear logistic  
6 regression can be attributed to its robustness when the assumptions underly-  
7 ing other classifiers are violated.

## 8 **References**

- 9 Bouchard, G., Triggs, B., 2004. The tradeoff between generative and discrim-  
10 inative classifiers. In: IASC International Symposium on Computational  
11 Statistics (COMPSTAT). Prague, pp. 721–728.
- 12 Dawid, A. P., 1976. Properties of diagnostic data distributions. *Biometrics*  
13 32 (3), 647–658.
- 14 Efron, B., 1975. The efficiency of logistic regression compared to normal dis-  
15 criminant analysis. *Journal of the American Statistical Association* 70 (352),  
16 892–898.
- 17 Fujino, A., Ueda, N., Saito, K., 2007. A hybrid generative/discriminative  
18 approach to text classification with additional information. *Information*  
19 *Processing and Management* 43 (2), 379–392.
- 20 McCallum, A., Pal, C., Druck, G., Wang, X., 2006. Multi-conditional learn-  
21 ing: Generative/discriminative training for clustering and classification. In:  
22 *AAAI*. pp. 433–439.
- 23 Ng, A. Y., Jordan, M. I., 2001. On discriminative vs. generative classifiers: a  
24 comparison of logistic regression and naïve bayes. In: *NIPS*.

- 1 Raina, R., Shen, Y., Ng, A. Y., McCallum, A., 2003. Classification with hybrid  
2 generative/discriminative models. In: NIPS.
- 3 Rubinstein, Y. D., Hastie, T., 1997. Discriminative vs. informative learning.  
4 In: KDD. pp. 49–53.
- 5 Titterington, D. M., Murray, G. D., Murray, L. S., Spiegelhalter, D. J., Skene,  
6 A. M., Habbema, J. D. F., Gelpke, G. J., 1981. Comparison of discrimi-  
7 nation techniques applied to a complex data set of head injured patients  
8 (with discussion). Journal of the Royal Statistical Society. Series A (Gen-  
9 eral) 144 (2), 145–175.