Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

The efficiency of logistic regression compared to normal discriminant analysis under class-conditional classification noise

Yingtao Bi, Daniel R. Jeske *

Department of Statistics, University of California, Riverside, CA 92521, United States

A R T I C L E I N F O

Article history:
Received 20 April 2009
Available online 18 March 2010

AMS subject classification:
62H30
62F12
62E20

Keywords:
Class noise
Misclassification rate
Misspecified model
Asymptotic distribution

A B S T R A C T

In many real world classification problems, class-conditional classification noise (CCC-Noise) frequently deteriorates the performance of a classifier that is naively built by ignoring it. In this paper, we investigate the impact of CCC-Noise on the quality of a popular generative classifier, normal discriminant analysis (NDA), and its corresponding discriminative classifier, logistic regression (LR). We consider the problem of two multivariate normal populations having a common covariance matrix. We compare the asymptotic distribution of the misclassification error rate of these two classifiers under CCC-Noise. We show that when the noise level is low, the asymptotic error rates of both procedures are only slightly affected. We also show that LR is less deteriorated by CCC-Noise compared to NDA. Under CCC-Noise contexts, the Mahalanobis distance between the populations plays a vital role in determining the relative performance of these two procedures. In particular, when this distance is small, LR tends to be more tolerable to CCC-Noise compared to NDA.

1. Introduction

In many real world classification problems, class label noise is inherent in the training dataset. CCC-Noise is an important type of class label noise. Examples where CCC-Noise arises include remote sensing and traditional medical diagnosis (see [1–3]) and image classification (see [4,5]). Recent medical diagnosis technologies (e.g. microarray analyses and protein mass-spectrometry profiling) have also sparked interest in considering classifiers where training data is prone to CCC-Noise (see, for example, [6,7] or [8]). In many applications, most analysts acknowledge that noise is present but naively ignore it. Alternatively, data preprocessing approaches have been proposed to attempt to remove mislabeled training observations, see [9] or [10]. These approaches, however, face the risk of removing useful data, which consequently would reduce the accuracy of the classifier. An alternative solution to handle noisy data might be to construct noise tolerant classifiers directly. Li et al. [4] combined the class noise into the model based on a probabilistic noise model. Norton and Hirsh [11] presented a Bayesian approach to learning from noisy data, where prior knowledge of the noise process is applied to compute posterior class probabilities. Yasui et al. [6] and Magder and Hughes [12] used EM algorithm to handle the label noise.

The effects of CCC-Noise on the estimation of association between two random variables have been discussed widely in epidemiology, see [13]. One important conclusion is that the noise attenuates the estimation of the association. Neuhaus [14] gave a comprehensive discussion of bias and efficiency loss due to CCC-Noise in the context of general linear model and proposed an approximation for the expected values of the estimators. The effects of CCC-Noise have also been studied in the area of pattern reorganization. Krishnan [15] studied the efficiency loss using Efron's Asymptotic Relative Efficiency criteria [16]. Michalek and Tripathi [3] bounded the asymptotic efficiency of logistic regression and normal discrimination.

* Corresponding author.
E-mail addresses: ybi001@ucr.edu (Y. Bi), daniel.jeske@ucr.edu (D.R. Jeske).

0047-259X/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2010.03.001

Zhu and Wu [17] investigated the impact of both class and attribute noise by doing simulation experiments. They concluded that the classification accuracies decline almost linearly with the increase of the noise level. Until now, limited research has been conducted to theoretically quantify the impact of CCC-Noise on the misclassification rates.

Generally speaking, classifiers can be characterized as either generative or discriminative, according to whether or not the distribution of the explanatory variables is modeled. Comparison of generative and discriminative classifiers has received considerable attention in the literature. There is a widely held perception that, in the absence of CCC-Noise, the discriminative approach is more robust than the generative approach. For example, Ng and Jordan [18] showed that if the assumed conditional distributions for the explanatory variables are correct, then generative classifiers and discriminative classifiers have the same asymptotic error rates, although the generative classifiers approach their asymptotic values faster. However, if the conditional distributions are not correct, then discriminative classifiers have lower asymptotic error rates provided the link function is modeled correctly. Efron [16] computed the relative efficiency, based on the ratio of expected regrets, of one popular generative classifier, normal discriminant analysis (NDA), to the corresponding discriminative classifier [18], logistic regression (LR), and concluded that LR is between one half and two thirds as efficient as NDA under normality. In this paper we aim at theoretically comparing the misclassification error rate of NDA with LR under CCC-Noise.

The rest of this paper is organized as follows. In Section 2 we briefly review the NDA and LR classifiers and then discuss these two procedures under CCC-Noise. In Section 3, the asymptotic distributions of the misclassification error rate of the NDA and LR under CCC-Noise are obtained. In Section 4, we compare the expected values of the misclassification error rates of these two procedures and also the relative efficiency. Section 5 concludes with summary.

2. NDA and LR classifiers under CCC-noise

We consider a binary classification task. Let **X** denote the vector of explanatory variables. Suppose that **X** comes from one of two p-dimensional normal populations differing in mean but not in covariance

$$\begin{array}{ll} \text{population } 0 : \mathbf{X} \sim \mathrm{MVN}_p(\boldsymbol{\mu_0}, \boldsymbol{\Sigma}) & \text{with probability } \pi_0, \\ \text{population } 1 : \mathbf{X} \sim \mathrm{MVN}_p(\boldsymbol{\mu_1}, \boldsymbol{\Sigma}) & \text{with probability } \pi_1, \end{array} \tag{1}$$

where $\pi_0 + \pi_1 = 1$.

If all parameters are known, a new observation may be classified based on **X** as belonging to population 1 or 0 according as

$$L_{(\alpha,\boldsymbol{\beta}')}(\mathbf{X}) = \alpha + \boldsymbol{\beta}'\mathbf{X} > \text{ or } < 0,$$

where

$$\begin{aligned} \alpha &= \lambda - (\boldsymbol{\mu}_1'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0)/2 \\ \boldsymbol{\beta} &= \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) \end{aligned} \tag{2}$$

with $\lambda = \log \pi_1/\pi_0$.

This linear classifier minimizes the total misclassification rate, as is easily shown by applying Bayes theorem. However, the assumption that the parameters $(\lambda, \boldsymbol{\mu_0}, \boldsymbol{\mu_1}, \boldsymbol{\Sigma})$ are known is rarely the case in practice. Therefore, the parameters are usually estimated from a random sample of n observations, $\{(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_n}, y_n)\}$ where y_k is the binary class label for the kth sample. Parameter estimates are typically obtained by maximizing the following log-likelihood function

$$\begin{aligned} &\sum_{k=1}^{n}(1-y_k)\left\{-\frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x_k}-\boldsymbol{\mu_0})'\boldsymbol{\Sigma}^{-1}(\mathbf{x_k}-\boldsymbol{\mu_0}) + \log\pi_0\right\} \\ &\quad + \sum_{k=1}^{n} y_k\left\{-\frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{x_k}-\boldsymbol{\mu_1})'\boldsymbol{\Sigma}^{-1}(\mathbf{x_k}-\boldsymbol{\mu_1}) + \log\pi_1\right\}. \end{aligned}$$

Using maximum likelihood estimators (MLEs) in place of unknown parameters results in the so called NDA procedure. The corresponding discriminative method to NDA is LR. LR is a popular discriminative method that does not assume any distribution for **X**. It simply assumes a parametric form for the conditional probability $P(Y = 1|\mathbf{X} = \mathbf{x})$. Then the parameters $(\alpha, \boldsymbol{\beta}')$ are estimated directly by maximizing the conditional likelihood

$$\sum_{k=1}^{n} y_k(\alpha + \boldsymbol{\beta}'\mathbf{x_k}) - \sum_{k=1}^{n}\log[1 + \exp(\alpha + \boldsymbol{\beta}'\mathbf{x_k})].$$

In the context of no label noise both NDA and LR will give consistent estimates under (1).

Now let's introduce the CCC-Noise into these two procedures. For a binary classification problem, CCC-Noise means that the label Y (0 or 1) of any observation is independently and randomly flipped to $1 - Y$ with probability $1 - \theta_Y$, where, θ_Y

denotes the correct labeling rate. We know that Y denotes the corresponding true class label, but instead of observing Y, we observe $\tilde{Y}$ as the noisy class label. The CCC-Noise can be specified as

$$P(\tilde{Y}=1|Y=0)=1-\theta_0$$
$$P(\tilde{Y}=0|Y=1)=1-\theta_1$$

and the observed training data is $\{(\mathbf{x_1},\tilde{y}_1),\dots,(\mathbf{x_n},\tilde{y}_n)\}$.

For the NDA procedure, the misspecified log-likelihood is

$$\sum_{k=1}^{n}(1-\tilde{y}_k)\left\{-\frac{p}{2}\log(2\pi)-\frac{1}{2}\log|\mathbf{\Sigma}|-\frac{1}{2}(\mathbf{x_k}-\boldsymbol{\mu_0})'\mathbf{\Sigma^{-1}}(\mathbf{x_k}-\boldsymbol{\mu_0})+\log\pi_0\right\}$$
$$+\sum_{k=1}^{n}\tilde{y}_k\left\{-\frac{p}{2}\log(2\pi)-\frac{1}{2}\log|\mathbf{\Sigma}|-\frac{1}{2}(\mathbf{x_k}-\boldsymbol{\mu_1})'\mathbf{\Sigma^{-1}}(\mathbf{x_k}-\boldsymbol{\mu_1})+\log\pi_1\right\}. \quad (3)$$

The MLEs obtained by maximizing (3) are termed misspecified MLEs. Similarly, the misspecified MLEs for the LR procedure are obtained by maximizing

$$\sum_{k=1}^{n}\tilde{y}_k(\alpha+\boldsymbol{\beta}'\mathbf{x_k})-\sum_{k=1}^{n}\log[1+\exp(\alpha+\boldsymbol{\beta}'\mathbf{x_k})]. \quad (4)$$

3. Asymptotic distribution of error rate

Let $(\tilde{\alpha},\tilde{\boldsymbol{\beta}}')$ denote an arbitrary estimate of $(\alpha,\boldsymbol{\beta}')$ based upon the training data set. For a new observation $(\mathbf{X},Y)$, the probability of a misclassification (calculated with respect to the joint distribution of the training data and the new observation) error is defined to be

$$\mathrm{ER}(\tilde{\alpha},\tilde{\boldsymbol{\beta}})=P(L_{(\tilde{\alpha},\tilde{\boldsymbol{\beta}})}(\mathbf{X})>0,Y=0)+P(L_{(\tilde{\alpha},\tilde{\boldsymbol{\beta}})}(\mathbf{X})<0,Y=1).$$

If $(\tilde{\alpha},\tilde{\boldsymbol{\beta}}')$ are randomly determined (as in NDA or LR), $\mathrm{ER}(\tilde{\alpha},\tilde{\boldsymbol{\beta}})$ will be a random variable. Efron [16] stated that the distribution of $\mathrm{ER}(\tilde{\alpha},\tilde{\boldsymbol{\beta}})$ is invariant under a linear transformation that allows reduction of assumption (1) to the following canonical form

$$\begin{array}{ll}\text{population } 0:\mathbf{X}\sim\mathrm{MVN}_p(-(\Delta/2)\mathbf{e_1},\mathbf{I}) & \text{with probability } \pi_0\\ \text{population } 1:\mathbf{X}\sim\mathrm{MVN}_p((\Delta/2)\mathbf{e_1},\mathbf{I}) & \text{with probability } \pi_1,\end{array} \quad (5)$$

where $\Delta\equiv\sqrt{(\boldsymbol{\mu_1}-\boldsymbol{\mu_0})'\mathbf{\Sigma^{-1}}(\boldsymbol{\mu_1}-\boldsymbol{\mu_0})}$, the square root of the Mahalanobis distance, $\mathbf{I}$ is the $p\times p$ identity matrix and $\mathbf{e}_1'\equiv(1,0,0,\dots,0)$ is a $1\times p$ vector. For either LR or NDA, the probability of misclassification has the same distribution under (1) as it does under the so-called "standard situation" (5). A formal proof of this statement is presented in [19] where it is also shown the result continues to hold even in the presence of CCC-Noise. Henceforth, we will work with the "standard situation".

In the standard situation, the conditional (given the training data) probability of misclassification of an arbitrary estimated classification boundary, $L_{(\tilde{\alpha},\tilde{\boldsymbol{\beta}})}=\tilde{\alpha}+\tilde{\boldsymbol{\beta}}'\mathbf{x}=0$, is

$$\mathrm{ER}(\tilde{\alpha},\tilde{\boldsymbol{\beta}})=\pi_0\Phi\left(\frac{-\frac{\Delta}{2}\tilde{\beta}_1+\tilde{\alpha}}{\sqrt{\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}}}}\right)+\pi_1\Phi\left(\frac{-\frac{\Delta}{2}\tilde{\beta}_1-\tilde{\alpha}}{\sqrt{\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}}}}\right), \quad (6)$$

where $\tilde{\beta}_1$ indicates the first component of $\tilde{\boldsymbol{\beta}}$ and $\Phi(.)$ is the cumulative density function of the standard normal distribution. Now, suppose the arbitrary estimate $(\tilde{\alpha},\tilde{\boldsymbol{\beta}}')$ has a limiting normal distribution

$$\sqrt{n}\left\{\begin{pmatrix}\tilde{\alpha}\\ \tilde{\boldsymbol{\beta}}\end{pmatrix}-\begin{pmatrix}\tilde{\alpha}^*\\ \tilde{\boldsymbol{\beta}}^*\end{pmatrix}\right\}\to\mathrm{MVN}_{p+1}\left[\mathbf{0},\begin{pmatrix}z_{00} & \mathbf{z}_{\mathbf{01}}'\\ \mathbf{z_{01}} & \mathbf{Z_{11}}\end{pmatrix}\right], \quad (7)$$

where z_{00} is a scalar, $\mathbf{z_{01}}$ is a column $p\times 1$ vector and $\mathbf{Z_{11}}$ is a $p\times p$ matrix. Differentiating (6) gives,

$$\frac{\partial\mathrm{ER}(\tilde{\alpha},\tilde{\boldsymbol{\beta}})}{\partial\tilde{\alpha}}=f_0(\tilde{\alpha},\tilde{\boldsymbol{\beta}})=\pi_0\frac{1}{\sqrt{\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}}}}\phi\left(\frac{-\frac{\Delta}{2}\tilde{\beta}_1+\tilde{\alpha}}{\sqrt{\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}}}}\right)-\pi_1\frac{1}{\sqrt{\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}}}}\phi\left(\frac{-\frac{\Delta}{2}\tilde{\beta}_1-\tilde{\alpha}}{\sqrt{\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}}}}\right),$$

$$\frac{\partial \mathrm{ER}(\tilde{\alpha},\tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}}} = \mathbf{f_1}(\tilde{\alpha},\tilde{\boldsymbol{\beta}}) = \pi_0\left[\frac{-\frac{\Delta}{2}}{\sqrt{\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}}}}\mathbf{e_1} + \left(\frac{\Delta}{2}\tilde{\beta}_1 - \tilde{\alpha}\right)(\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}})^{-\frac{3}{2}}\tilde{\boldsymbol{\beta}}\right]\phi\left(\frac{-\frac{\Delta}{2}\tilde{\beta}_1 + \tilde{\alpha}}{\sqrt{\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}}}}\right)$$
$$+\pi_1\left[\frac{-\frac{\Delta}{2}}{\sqrt{\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}}}}\mathbf{e_1} + \left(\frac{\Delta}{2}\tilde{\beta}_1 + \tilde{\alpha}\right)(\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}})^{-\frac{3}{2}}\tilde{\boldsymbol{\beta}}\right]\phi\left(\frac{-\frac{\Delta}{2}\tilde{\beta}_1 - \tilde{\alpha}}{\sqrt{\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}}}}\right),$$

where ϕ is the standard normal density function.

Defining $\omega^2(\tilde{\alpha}^*,\tilde{\boldsymbol{\beta}}^*) = f_0^2(\tilde{\alpha}^*,\tilde{\boldsymbol{\beta}}^*)z_{00} + \mathbf{f}_1'(\tilde{\alpha}^*,\tilde{\boldsymbol{\beta}}^*)\mathbf{Z_{11}}\mathbf{f_1}(\tilde{\alpha}^*,\tilde{\boldsymbol{\beta}}^*) + 2f_0(\tilde{\alpha}^*,\tilde{\boldsymbol{\beta}}^*)\mathbf{z}_{\mathbf{01}}'\mathbf{f}_1(\tilde{\alpha}^*,\tilde{\boldsymbol{\beta}}^*)$, the following lemma follows from application of the delta method:

Lemma 1. *In the standard situation, if* $(\tilde{\alpha},\tilde{\boldsymbol{\beta}}')$ *has the limiting distribution given in* (7) *with*

$$(\tilde{\alpha}^*,\tilde{\boldsymbol{\beta}}^{*\prime}) \neq (\alpha,\boldsymbol{\beta}'), \quad \textit{then}\ \sqrt{n}[\mathrm{ER}(\tilde{\alpha},\tilde{\boldsymbol{\beta}}) - \mathrm{ER}(\tilde{\alpha}^*,\tilde{\boldsymbol{\beta}}^*)] \xrightarrow{L} N(0,\omega^2(\tilde{\alpha}^*,\tilde{\boldsymbol{\beta}}^*)).$$

Note that if $(\tilde{\alpha}^*,\tilde{\boldsymbol{\beta}}^{*\prime})$ are the true values of $(\alpha,\boldsymbol{\beta}')$, then $f_0(\tilde{\alpha}^*,\tilde{\boldsymbol{\beta}}^*)$ and $\mathbf{f_1}(\tilde{\alpha}^*,\tilde{\boldsymbol{\beta}}^*)$ will both vanish since $\mathrm{ER}(\tilde{\alpha},\tilde{\boldsymbol{\beta}})$ is minimized at that point. In this situation, Lemma 1 will not be valid since the first order Taylor expansion is not enough. Efron [16] computed the efficiency of LR to NDA, without CCC-Noise, based on the second order Taylor expansion. For the case of having CCC-Noise, Lemma 1 will be valid since $(\tilde{\alpha}^*,\tilde{\boldsymbol{\beta}}^{*\prime})$ are not equal to the true values. In order to use Lemma 1 to obtain the asymptotic distribution of $\mathrm{ER}(\tilde{\alpha},\tilde{\boldsymbol{\beta}})$, we will first to get the asymptotic distribution of $(\tilde{\alpha},\tilde{\boldsymbol{\beta}}')$. To this end, we will utilize White's theorem [20].

White's Theorem. *Suppose i.i.d. data* $\mathbf{x_1},\mathbf{x_2},\ldots,\mathbf{x_n}$ *come from a true distribution* $p(\mathbf{x})$*, but we assume a family of distributions,* $p(\mathbf{x}|\boldsymbol{\theta})$ *where* $\boldsymbol{\theta}$ *is an indexing parameter. Then under suitable regularity conditions, the maximum likelihood estimator of* $\boldsymbol{\theta}$ *converges almost surely to the value* $\boldsymbol{\theta}^*$ *that minimizes* $-\int p(\mathbf{x})\log\left[\frac{p(\mathbf{x}|\boldsymbol{\theta})}{p(\mathbf{x})}\right]\mathrm{d}\mathbf{x}$*, the Kullback–Leibler distance of* $p(\mathbf{x}|\boldsymbol{\theta})$ *from* $p(\mathbf{x})$*. White also shows that the sequence of MLEs,* $\hat{\boldsymbol{\theta}}_{\mathbf{n}}$*, is asymptotically multivariate normal in the sense* $\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathbf{n}} - \boldsymbol{\theta}^*) \xrightarrow{L}$ $\mathrm{MVN}(\mathbf{0},\mathbf{C}(\boldsymbol{\theta}^*))$*, where the covariance matrix* $\mathbf{C}(\boldsymbol{\theta}^*)$ *is equal to* $\mathbf{A}(\boldsymbol{\theta}^*)^{-\mathbf{1}}\mathbf{B}(\boldsymbol{\theta}^*)\mathbf{A}(\boldsymbol{\theta}^*)^{-\mathbf{1}}$*, with* $\mathbf{A}(\boldsymbol{\theta})$ *and* $\mathbf{B}(\boldsymbol{\theta})$ *matrices whose* (i,j)*th element is given by*

$$\mathbf{A_{ij}}(\theta) = E_{\mathbf{p(x)}}[\partial^2 \log p(\mathbf{x}|\boldsymbol{\theta})/\partial\theta_i\theta_j],$$
$$\mathbf{B_{ij}}(\theta) = E_{\mathbf{p(x)}}[(\partial \log p(\mathbf{x}|\boldsymbol{\theta})/\partial\theta_i)(\partial \log p(\mathbf{x}|\boldsymbol{\theta})/\partial\theta_j)].$$

3.1. Asymptotic distribution of error rate of NDA

First we consider the misspecified NDA procedure. Let $(\hat{\alpha},\hat{\boldsymbol{\beta}}')$ and $(\hat{\lambda},\hat{\boldsymbol{\mu}}_{\mathbf{0}},\hat{\boldsymbol{\mu}}_{\mathbf{1}},\hat{\boldsymbol{\Sigma}})$ denote, respectively, the estimates of $(\alpha,\boldsymbol{\beta}')$ and $(\lambda,\boldsymbol{\mu}_{\mathbf{0}},\boldsymbol{\mu}_{\mathbf{1}},\boldsymbol{\Sigma})$ based on the misspecified NDA procedure. Let us write the distinct elements of $\hat{\boldsymbol{\Sigma}}^{-\mathbf{1}}$ as a $p(p+1)/2$ vector $(\hat{\sigma}^{11},\hat{\sigma}^{12},\ldots,\hat{\sigma}^{1p},\hat{\sigma}^{22},\hat{\sigma}^{23},\ldots,\hat{\sigma}^{pp})$ and indicate this vector as $(\hat{\boldsymbol{\sigma}}^{(1)},\hat{\boldsymbol{\sigma}}^{(2)})$, where $\hat{\boldsymbol{\sigma}}^{(1)} \equiv (\hat{\sigma}^{11},\hat{\sigma}^{12},\ldots,\hat{\sigma}^{1p})$, $\hat{\boldsymbol{\sigma}}^{(2)} \equiv (\hat{\sigma}^{22},\hat{\sigma}^{23},\ldots,\hat{\sigma}^{pp})$.

For convenience we introduce the following expressions first,

$$a = \pi_0\theta_0, b = \pi_1(1-\theta_1), c = \pi_0(1-\theta_0), d = \pi_1\theta_1, \quad h = 1 + \left(\frac{ab}{a+b} + \frac{cd}{c+d}\right)\Delta^2.$$

We also have the following notations, $\mathbf{E_{ij}}$ being the $p\times p$ matrix with one in the (i,j)th position and zero elsewhere, $\mathbf{O_{p\times p}}$ being zero matrixes, and $\delta_{ij} = 1$ or 0 as $i = j$ or $i \neq j$. As described earlier, the asymptotic distribution of the error rate is the same under (1) as it is under (5), so we consider the asymptotic distribution of the misspecified MLEs in the standard situation. It will also be seen that we do not need the limiting distribution of $\hat{\boldsymbol{\sigma}}^{(2)}$ to derive the asymptotic distribution of NDA based estimates $(\hat{\alpha},\hat{\boldsymbol{\beta}}')$.

Lemma 2. *In the standard situation, under CCC-Noise, the NDA produces misspecified estimates* $(\hat{\lambda},\hat{\boldsymbol{\mu}}_{\mathbf{0}},\hat{\boldsymbol{\mu}}_{\mathbf{1}},\hat{\boldsymbol{\sigma}}^{(1)})$*, by maximizing* (3)*, which satisfy*

$$\sqrt{n}\left\{\begin{pmatrix}\hat{\lambda}\\ \hat{\boldsymbol{\mu}}_{\mathbf{0}}\\ \hat{\boldsymbol{\mu}}_{\mathbf{1}}\\ \hat{\boldsymbol{\sigma}}^{(1)}\end{pmatrix} - \begin{pmatrix}\lambda^*\\ \boldsymbol{\mu}_{\mathbf{0}}^*\\ \boldsymbol{\mu}_{\mathbf{1}}^*\\ \boldsymbol{\sigma}^{*(1)}\end{pmatrix}\right\} \to \mathrm{MVN}_{3p+1}(\mathbf{0},\boldsymbol{\Omega})$$

where

$$\lambda^* = \log\frac{c+d}{a+b}, \qquad \boldsymbol{\mu}_\mathbf{0}^* = \frac{b-a}{b+a}\frac{\Delta}{2}\mathbf{e_1}, \qquad \boldsymbol{\mu}_\mathbf{1}^* = \frac{d-c}{c+d}\frac{\Delta}{2}\mathbf{e_1}, \qquad \boldsymbol{\sigma}^{(1)*} = \mathbf{e_1}/h$$

and

$$\boldsymbol{\Omega} = \begin{pmatrix} \omega_{00} & \mathbf{O_{1\times p}} & \mathbf{O_{1\times p}} & \omega_{03}\mathbf{e}_\mathbf{1}' \\ \mathbf{O_{p\times 1}} & \omega_{11}\mathbf{E_{11}} + \dfrac{\mathbf{I}-\mathbf{E_{11}}}{a+b} & \mathbf{O_{p\times p}} & \omega_{13}\mathbf{E_{11}} \\ \mathbf{O_{p\times 1}} & \mathbf{O_{p\times p}} & \omega_{22}\mathbf{E_{11}} + \dfrac{\mathbf{I}-\mathbf{E_{11}}}{c+d} & \omega_{23}\mathbf{E_{11}} \\ \omega_{03}\mathbf{e_1} & \omega_{13}\mathbf{E_{11}} & \omega_{23}\mathbf{E_{11}} & \omega_{33}\mathbf{E_{11}} + h(\mathbf{I}-\mathbf{E_{11}}) \end{pmatrix}$$

with

$$\omega_{00} = \frac{[1+\exp(\lambda^*)]^2}{\exp(\lambda^*)} = \frac{1}{(a+b)(c+d)}, \qquad \omega_{11} = \frac{1}{a+b} + \frac{ab}{(a+b)^3}\Delta^2,$$

$$\omega_{22} = \frac{1}{c+d} + \frac{ab}{(c+d)^3}\Delta^2,$$

$$\omega_{33} = \frac{2}{h^2} + \frac{\Delta^4}{h^4}\left[\frac{ab^4+a^4b}{(b+a)^4} + \frac{cd^4+c^4d}{(d+c)^4} - 3\left(\frac{ab}{b+a} + \frac{cd}{d+c}\right)^2\right],$$

$$\omega_{03} = \frac{\Delta^2}{h^2}\left[\frac{ab}{(a+b)^2} - \frac{cd}{(c+d)^2}\right], \qquad \omega_{13} = \frac{ab(a-b)\Delta^3}{(a+b)^3h^2}, \qquad \omega_{23} = \frac{cd(c-d)\Delta^3}{(c+d)^3h^2}.$$

Proof. see Appendix A. □

Lemma 2 gives us the asymptotic distribution of the misspecified MLEs $\hat{\lambda}$, $\hat{\boldsymbol{\mu}}_\mathbf{0}$, $\hat{\boldsymbol{\mu}}_\mathbf{1}$ and $\hat{\boldsymbol{\sigma}}^{(1)}$. Similar to Efron's derivations [16], we use the multivariate delta method to obtain Lemma 3 which uses the following definitions,

$$m_{01} = \frac{b-a}{b+a}\frac{\Delta}{2h}, \qquad m_{02} = \frac{c-d}{c+d}\frac{\Delta}{2h},$$

$$m_{03} = \left[\left(\frac{b-a}{b+a}\right)^2 - \left(\frac{d-c}{d+c}\right)^2\right]\frac{\Delta^2}{8}, \qquad m_{13} = \left(\frac{d-c}{d+c} - \frac{b-a}{b+a}\right)\frac{\Delta}{2}.$$

Lemma 3. *In the standard situation, under CCC-Noise, the NDA produces misspecified estimates* $(\hat{\alpha}, \hat{\boldsymbol{\beta}}')$ *satisfying*

$$\sqrt{n}\left\{\begin{pmatrix}\hat{\alpha}\\ \hat{\boldsymbol{\beta}}\end{pmatrix} - \begin{pmatrix}\alpha^*\\ \boldsymbol{\beta}^*\end{pmatrix}\right\} \to \mathrm{MVN}_{p+1}(\mathbf{0}, \boldsymbol{\Lambda}),$$

where

$$\alpha^* = \log\frac{c+d}{a+b} - \frac{1}{2h}\left[\left(\frac{d-c}{d+c}\right)^2 - \left(\frac{b-a}{b+a}\right)^2\right]\left(\frac{\Delta}{2}\right)^2,$$

$$\boldsymbol{\beta}^* = (\boldsymbol{\mu}_\mathbf{1}^* - \boldsymbol{\mu}_\mathbf{0}^*)(\boldsymbol{\Sigma}^*)^{-1} = \frac{\Delta}{h}\frac{ad-bc}{(b+a)(c+d)}\mathbf{e_1},$$

$$\boldsymbol{\Lambda} = \begin{pmatrix} \kappa_{00} & \kappa_{01}\mathbf{e}_\mathbf{1}' \\ \kappa_{01}\mathbf{e_1} & \kappa_{11}\mathbf{E_{11}} + \left[\dfrac{1}{(a+b)(c+d)h^2} + m_{13}^2h\right](\mathbf{I}-\mathbf{E_{11}}) \end{pmatrix}$$

with

$$\kappa_{00} = \omega_{00} + \omega_{11}(m_{01})^2 + \omega_{22}(m_{02})^2 + \omega_{33}(m_{03})^2 + 2m_{03}\omega_{03} + 2m_{03}\omega_{13}m_{01} + 2m_{03}\omega_{23}m_{02},$$

$$\begin{aligned}\kappa_{01} \;=\; & \frac{-m_{01}\omega_{11}}{h} + \frac{m_{02}\omega_{22}}{h} + m_{03}\omega_{33}m_{13} + \omega_{03}m_{13} + \frac{-m_{03}\omega_{13}}{h} \\ & + m_{01}\omega_{13}m_{13} + \frac{m_{03}\omega_{23}}{h} + m_{02}\omega_{23}m_{13},\end{aligned}$$

$$\kappa_{11} = \frac{\omega_{11}+\omega_{22}}{h^2} + m_{13}^2\omega_{33} - \frac{2m_{13}\Delta^3}{h^3}\left[\frac{ab(a-b)}{(a+b)^3} - \frac{cd(c-d)}{(c+d)^3}\right].$$

Proof. Differentiating (2) gives

$$\frac{\partial\alpha}{\partial\lambda}=1,\qquad \frac{\partial\alpha}{\partial\boldsymbol{\mu}_{\mathbf{0}}'}=\boldsymbol{\mu}_{\mathbf{0}}'\boldsymbol{\Sigma}^{\mathbf{-1}},\qquad \frac{\partial\alpha}{\partial\boldsymbol{\mu}_1'}=-\boldsymbol{\mu}_{\mathbf{1}}'\boldsymbol{\Sigma}^{\mathbf{-1}},\qquad \frac{\partial\alpha}{\partial\sigma^{ij}}=\frac{\mu_{0i}\mu_{0j}-\mu_{1i}\mu_{1j}}{1+\delta_{ij}},$$
$$\frac{\partial\boldsymbol{\beta}}{\partial\lambda}=\mathbf{0},\qquad \frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{\mu}_0'}=-\boldsymbol{\Sigma}^{\mathbf{-1}},\qquad \frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{\mu}_1'}=\boldsymbol{\Sigma}^{\mathbf{-1}},\qquad \frac{\partial\boldsymbol{\beta}}{\partial\sigma^{ij}}=\frac{\mathbf{E_{ij}}+\mathbf{E_{ji}}}{1+\delta_{ij}}(\boldsymbol{\mu}_{\mathbf{1}}-\boldsymbol{\mu}_{\mathbf{0}}).$$

μ_{0i} indicating the ith component of $\boldsymbol{\mu}_{\mathbf{0}}$; likewise for μ_{1i}. In the standard situation, we have the following first-order differential relationship that is obtained by expanding $(\hat{\alpha},\hat{\boldsymbol{\beta}}')$ around $(\alpha^*,\boldsymbol{\beta}^{*\prime})$

$$\begin{pmatrix}\hat{\alpha}\\ \hat{\boldsymbol{\beta}}\end{pmatrix}-\begin{pmatrix}\alpha^*\\ \boldsymbol{\beta}^*\end{pmatrix}=\begin{bmatrix}1 & m_{01}\mathbf{e}'{}_{\mathbf{1}} & m_{02}\mathbf{e}_{\mathbf{1}}' & m_{03}\mathbf{e}_{\mathbf{1}}' & \mathbf{0}\\ \mathbf{0} & -(\boldsymbol{\Sigma}^*)^{-1} & (\boldsymbol{\Sigma}^*)^{-1} & m_{13}\mathbf{I} & \mathbf{0}\end{bmatrix}\begin{pmatrix}\hat{\lambda}-\lambda^*\\ \hat{\boldsymbol{\mu}}_{\mathbf{0}}-\boldsymbol{\mu}_{\mathbf{0}}^*\\ \hat{\boldsymbol{\mu}}_{\mathbf{1}}-\boldsymbol{\mu}_{\mathbf{1}}^*\\ \hat{\boldsymbol{\sigma}}^{(1)}-\boldsymbol{\sigma}^{*(1)}\\ \hat{\boldsymbol{\sigma}}^{(2)}-\boldsymbol{\sigma}^{*(2)}\end{pmatrix}. \tag{8}$$

Let $\mathbf{M}$ be the matrix on the right side of (8), ignoring the last column. Then the multivariate delta method gives the covariance matrix of $(\hat{\alpha},\hat{\boldsymbol{\beta}}')$ as $\mathbf{M\Omega M}'$. Evaluation of $\mathbf{M\Omega M}'$ gives the result of $\boldsymbol{\Lambda}$. □

Then the asymptotic distribution of $\mathrm{ER}(\hat{\alpha},\hat{\boldsymbol{\beta}})$ is given in Theorem 1.

Theorem 1.

$$\sqrt{n}[\mathrm{ER}(\hat{\alpha},\hat{\boldsymbol{\beta}})-\mathrm{ER}(\alpha^*,\beta_1^*\mathbf{e_1})]\xrightarrow{L}N(0,\omega_{\mathrm{NDA}}^2),$$

where

$$\mathrm{ER}(\alpha^*,\beta_1^*\mathbf{e_1})=\pi_0\Phi\left(-\frac{\Delta}{2}+\frac{\alpha^*}{|\beta_1^*|}\right)+\pi_1\Phi\left(-\frac{\Delta}{2}-\frac{\alpha^*}{|\beta_1^*|}\right),$$
$$\omega_{\mathrm{NDA}}^2=(f_0(\alpha^*,\beta_1^*\mathbf{e_1}))^2\kappa_{00}+\kappa_{11}(f_{11}(\alpha^*,\beta_1^*\mathbf{e_1}))^2+2f_0(\alpha^*,\beta_1^*\mathbf{e_1})\kappa_{01}f_{11}(\alpha^*,\beta_1^*\mathbf{e_1}).$$

with

$$f_0(\alpha^*,\beta_1^*\mathbf{e_1})=\frac{\pi_0}{|\beta_1^*|}\phi\left(-\frac{\Delta}{2}+\frac{\alpha^*}{|\beta_1^*|}\right)-\frac{\pi_1}{\beta_1^*}\phi\left(-\frac{\Delta}{2}-\frac{\alpha^*}{|\beta_1^*|}\right),$$
$$f_{11}(\alpha^*,\beta_1^*\mathbf{e_1})=-\pi_0\frac{\alpha^*}{(\beta_1^*)^2}\phi\left(-\frac{\Delta}{2}+\frac{\alpha^*}{|\beta_1^*|}\right)+\pi_1\frac{\alpha^*}{(\beta_1^*)^2}\phi\left(-\frac{\Delta}{2}-\frac{\alpha^*}{|\beta_1^*|}\right).$$

Proof. The proof follows immediately from Lemmas 1 and 3. □

We know that the minimal value of the error rate is $\mathrm{ER}(\lambda,\Delta\mathbf{e_1})=\pi_0\Phi\left(-\frac{\Delta}{2}+\frac{\lambda}{\Delta}\right)+\pi_1\Phi\left(-\frac{\Delta}{2}-\frac{\lambda}{\Delta}\right)$. That means the misspecified NDA procedure will converge to $\mathrm{ER}(\lambda,\Delta\mathbf{e_1})$ if the following condition is satisfied

$$\frac{\alpha^*}{|\beta_1^*|}=\frac{\log\frac{c+d}{a+b}-\frac{1}{2h}\left[\left(\frac{d-c}{d+c}\right)^2-\left(\frac{b-a}{b+a}\right)^2\right]\left(\frac{\Delta}{2}\right)^2}{\frac{\Delta}{h}\frac{|ad-bc|}{(b+a)(c+d)}}=\frac{\lambda}{\Delta}. \tag{9}$$

When $\pi_0=\pi_1$, it can be easily seen that this equation holds if $\theta_0=\theta_1$.

3.2. Asymptotic distribution of error rate of LR

According to White's theorem, the misspecified MLEs, obtained by maximizing (4), say $(\hat{s},\hat{\mathbf{t}}')$ are asymptotically normal and converge almost surely to the value $(s^*,\mathbf{t}^{*\prime})$ that minimizes

$$E_{(\mathbf{X},\tilde{Y})}\left\{-\tilde{Y}(\alpha+\boldsymbol{\beta}'\mathbf{X})+\log\left[1+\exp(\alpha+\boldsymbol{\beta}'X)\right]-\log[\varphi(\mathbf{X})]\right\}, \tag{10}$$

where $\varphi(\mathbf{X})$ is the density function of $\mathbf{X}$ under (5).

Taking the first partial derivatives with respect to α and $\boldsymbol{\beta}$ respectively and setting them to zero gives:

$$E_{(\mathbf{X},\tilde{Y})}\left[\tilde{Y}\right]-E_{(\mathbf{X},\tilde{Y})}\left[\frac{\exp(s^*)\exp(\mathbf{t}^{*\prime}\mathbf{X})}{1+\exp(s^*)\exp(\mathbf{t}^{*\prime}\mathbf{X})}\right]=0, \tag{11}$$

$$E_{(\mathbf{X},\tilde{Y})}[\mathbf{X}\tilde{Y}]-E_{(\mathbf{X},\tilde{Y})}\left[\mathbf{X}\frac{\exp(s^*)\exp(\mathbf{t}^{*\prime}\mathbf{X})}{1+\exp(s^*)\exp(\mathbf{t}^{*\prime}\mathbf{X})}\right]=0. \tag{12}$$

Neuhaus [14] showed that ignoring CCC-Noise in the LR can be viewed as one kind of link function violation. Li and Duan's [21] showed that the estimated slope $\hat{\mathbf{t}}'$ from LR with a misspecified link function consistently estimates the true parameter $\boldsymbol{\beta}'$ up to a scale factor. In the standard situation $\boldsymbol{\beta}' = \Delta\mathbf{e}_1'$, thus we have $\mathbf{t}^{*\prime} = u^*\Delta\mathbf{e}_1'$, where u^* is a scalar quantity. Then (11) and (12) are simplified into,

$$g_1(\pi_0,\Delta,\theta_0,\theta_1) = E_{X_1}\left[\frac{\exp(s^*)\exp(u^*\Delta X_1)}{1+\exp(s^*)\exp(u^*\Delta X_1)}\right], \tag{13}$$

$$g_2(\pi_0,\Delta,\theta_0,\theta_1) = E_{X_1}\left[\frac{X_1\exp(s^*)\exp(u^*\Delta X_1)}{1+\exp(s^*)\exp(u^*\Delta X_1)}\right] \tag{14}$$

with

$$g_1(\pi_0,\Delta,\theta_0,\theta_1) = E_{(\mathbf{X},\tilde{Y})}\left[\tilde{Y}\right] = \theta_1 E_{X_1}\left[\frac{\pi_1\exp(\Delta X_1)}{\pi_0+\pi_1\exp(\Delta X_1)}\right] + (1-\theta_0)E_{X_1}\left[\frac{\pi_0}{\pi_0+\pi_1\exp(\Delta X_1)}\right],$$

$$g_2(\pi_0,\Delta,\theta_0,\theta_1) = E_{(\mathbf{X},\tilde{Y})}\left[\tilde{Y}X_1\right] = \theta_1 E_{X_1}\left[\frac{X_1\pi_1\exp(\Delta X_1)}{\pi_0+\pi_1\exp(\Delta X_1)}\right] + (1-\theta_0)E_{X_1}\left[\frac{X_1\pi_0}{\pi_0+\pi_1\exp(\Delta X_1)}\right]$$

and X_1 indicating the first component of $\mathbf{X}$. Eqs. (13) and (14) are solved by combining Monte-Carlo integration with the Newton–Raphson algorithm. The two unknown parameters (s^*, u^*) only depend on the values of $(\pi_0, \Delta, \theta_0, \theta_1)$. Instead of using $s^*(\pi_0,\Delta,\theta_0,\theta_1)$ and $u^*(\pi_0,\Delta,\theta_0,\theta_1)$, we adopt (s^*, u^*) for the convenience of expression.

Lemma 4. *In the standard situation, under CCC-Noise, the LR produces misspecified estimates $(\hat{s}, \hat{\mathbf{t}}')$ satisfying*

$$\sqrt{n}\left\{\begin{pmatrix}\hat{s}\\ \hat{\mathbf{t}}\end{pmatrix} - \begin{pmatrix}s^*\\ u^*\Delta\mathbf{e}_1\end{pmatrix}\right\} \to \mathrm{MVN}_p(\mathbf{0}, \boldsymbol{\Psi}),$$

where (s^*, u^*) is solution of (13) and (14), and

$$\boldsymbol{\Psi} = \begin{pmatrix}\psi_{00} & \psi_{01}\mathbf{e}'_1\\ \psi_{01}\mathbf{e}_1 & \psi_{11}\mathbf{E}_{11} + \dfrac{v_0}{h_0^2}(\mathbf{I}-\mathbf{E}_{11})\end{pmatrix}$$

with

$$\psi_{00} = \frac{h_2^2v_0 - 2h_1h_2v_1 + h_1^2v_2}{(h_0h_2-h_1^2)^2}, \qquad \psi_{01} = \frac{-h_1h_2v_0 + (h_1h_0+h_1^2)v_1 - h_0h_1v_2}{(h_0h_2-h_1^2)^2},$$

$$\psi_{11} = \frac{h_1^2v_0 - 2h_1h_0v_1 + h_0^2v_2}{(h_0h_2-h_1^2)^2}, \qquad h_i(\pi_0,\Delta,\theta_0,\theta_1) \equiv \int_{-\infty}^{+\infty}\frac{\exp(s^*)\exp(u^*\Delta x)}{[1+\exp(s^*)\exp(u^*\Delta x)]^2}x^i\varphi(x)\mathrm{d}x,$$

$$v_i(\pi_0,\Delta,\theta_0,\theta_1) \equiv \int_{-\infty}^{+\infty}\left\{\frac{\pi_1\theta_1\exp(\Delta x)+\pi_0(1-\theta_0)}{\pi_0+\pi_1\exp(\Delta x)}\,\frac{1-\exp(s^*)\exp(u^*\Delta x)}{1+\exp(s^*)\exp(u^*\Delta x)} + \left[\frac{\exp(s^*)\exp(u^*\Delta x)}{1+\exp(s^*)\exp(u^*\Delta x)}\right]\right\}x^i\varphi(x)\mathrm{d}x \quad i=0,1,2.$$

Proof. see Appendix B. □

Combining Lemmas 1 and 4 proves the following theorem regarding the asymptotic distribution of $\mathrm{ER}(\hat{s}, \hat{\mathbf{t}})$.

Theorem 2.

$$\sqrt{n}[\mathrm{ER}(\hat{s},\hat{\mathbf{t}}) - \mathrm{ER}(s^*, u^*\Delta\mathbf{e}_1)] \xrightarrow{L} N(0, \omega_{LR}^2),$$

where

$$\mathrm{ER}(s^*, u^*\Delta\mathbf{e}_1) = \pi_0\Phi\left(-\frac{\Delta}{2}+\frac{s^*}{|u^*|\Delta}\right) + \pi_1\Phi\left(-\frac{\Delta}{2}-\frac{s^*}{|u^*|\Delta}\right),$$

$$\omega_{LR}^2 = (f_0(s^*, u^*\Delta\mathbf{e}_1))^2\psi_{00} + \psi_{11}(f_{11}(s^*, u^*\Delta\mathbf{e}_1))^2 + 2f_0(s^*, u^*\Delta\mathbf{e}_1)\psi_{01}f_{11}(s^*, u^*\Delta\mathbf{e}_1)$$

with

$$f_0(s^*, u^*\Delta\mathbf{e}_1) = \pi_0\frac{1}{|u^*|\Delta}\phi\left(-\frac{\Delta}{2}+\frac{s^*}{|u^*|\Delta}\right) - \pi_1\frac{1}{|u^*|\Delta}\phi\left(-\frac{\Delta}{2}-\frac{s^*}{|u^*|\Delta}\right),$$

$$f_{11}(s^*, u^*\Delta\mathbf{e}_1) = -\pi_0\frac{s^*}{(u^*\Delta)^2}\phi\left(-\frac{\Delta}{2}+\frac{s^*}{|u^*|\Delta}\right) + \pi_1\frac{s^*}{(u^*\Delta)^2}\phi\left(-\frac{\Delta}{2}-\frac{s^*}{|u^*|\Delta}\right).$$

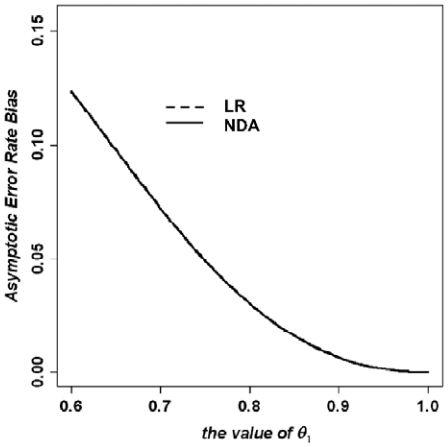

Fig. 1a. The asymptotic error rate bias with $\theta_0 = 1$, $\Delta = 1$ and $\pi_0 = 0.5$.

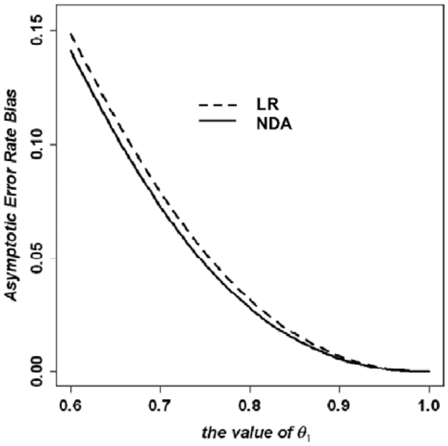

Fig. 1b. The asymptotic error rate bias with $\theta_0 = 1$, $\Delta = 2$ and $\pi_0 = 0.5$.

4. Comparison and discussion

Define the bias of the asymptotic error rate of the classification rule by

$$AERB(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = \lim_{n\to\infty} E\left[\text{ER}(\hat{\alpha}, \hat{\boldsymbol{\beta}})\right] - \text{ER}(\lambda, \Delta\mathbf{e_1}) = \text{ER}(\alpha^*, \beta_1^*\mathbf{e_1}) - \text{ER}(\lambda, \Delta\mathbf{e_1}),$$

$$AERB(\hat{s}, \hat{\mathbf{t}}) = \lim_{n\to\infty} E\left[\text{ER}(\hat{s}, \hat{\mathbf{t}})\right] - \text{ER}(\lambda, \Delta\mathbf{e_1}) = \text{ER}(s^*, u^*\Delta\mathbf{e_1}) - \text{ER}(\lambda, \Delta\mathbf{e_1}).$$

We assume that both θ_0 and θ_1 are greater than 0.5 since values of θ_0 and θ_1 less than 0.5 indicates that the process of labeling an observation performs worse than flipping a coin. Figs. 1–3 illustrate how the CCC-Noise affects the bias of the asymptotic error rate of NDA and LR respectively. If the CCC-Noise is ignored, both the LR and NDA procedures are positively biased. However, when the noise level is low, which is usually the case in practice, the asymptotic error rates of both procedures are only slightly affected. When Δ is small, the LR and NDA almost have the same asymptotic error rate. As Δ increases, the difference in the asymptotic error rates increases with LR always being larger. When $\theta_0 = \theta_1$, *AERB* is zero for both NDA and LR, implying the corresponding asymptotic error rates are both equal to the optimal value given by $\text{ER}(\lambda, \Delta\mathbf{e_1})$. We define the relative efficiency of LR to NDA as the following,

$$\text{RE}(\pi_0, \Delta, \theta_0, \theta_1, n) = \frac{E\left[\text{ER}(\hat{\alpha}, \hat{\boldsymbol{\beta}}) - \text{ER}(\lambda, \Delta\mathbf{e_1})\right]^2}{E\left[\text{ER}(\hat{s}, \hat{\mathbf{t}}) - \text{ER}(\lambda, \Delta\mathbf{e_1})\right]^2} \approx \frac{AERB^2(\hat{\alpha}, \hat{\boldsymbol{\beta}}) + \omega^2_{\text{NDA}}/n}{AERB^2(\hat{s}, \hat{\mathbf{t}}) + \omega^2_{\text{LR}}/n}. \tag{15}$$

The quantities of (15) are graphed in Figs. 4–6 for different sample sizes. These three graphs show that the relative efficiency monotonically increases as the noise level increases, which indicates that LR is less deteriorated by CCC-Noise compared to NDA. As noise level goes to zero, the relative efficiencies converge to their minimal values, which are different from the numbers shown in Table 1 of Efron's paper [16], because the relative efficiency employed in Efron's paper is defined as the

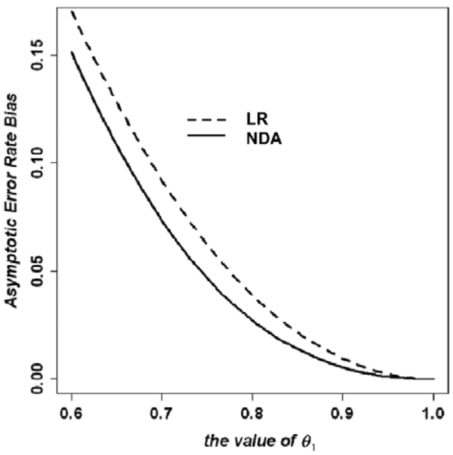

Fig. 1c. The asymptotic error rate bias with $\theta_0 = 1$, $\Delta = 3$ and $\pi_0 = 0.5$.

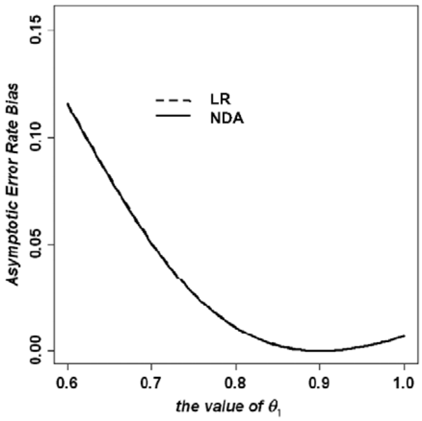

Fig. 2a. The asymptotic error rate bias with $\theta_0 = 0.9$, $\Delta = 1$ and $\pi_0 = 0.5$.

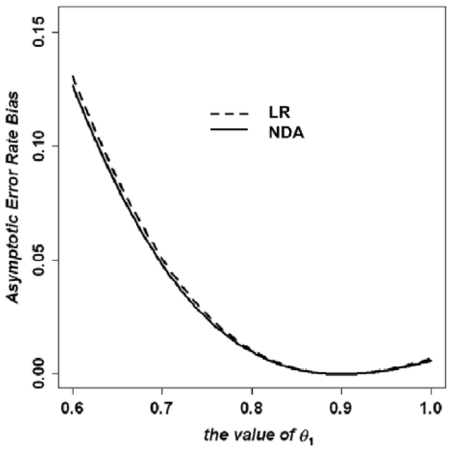

Fig. 2b. The asymptotic error rate bias with $\theta_0 = 0.9$, $\Delta = 2$ and $\pi_0 = 0.5$.

ratio of the expected regrets of these two procedures, rather than the ratio of the mean square errors as in our paper. It is also important to note that as Δ gets smaller, although NDA is still better it is better by a small margin. Furthermore, for the case of sample size equal to 50 and Δ equal to 2, as misclassification probability increases, the performance of LR can even be better than NDA. Though only Fig. 4 shows this latter aspect, for lower values of θ_1 some of these other curves would rise up to be larger than 1 also.

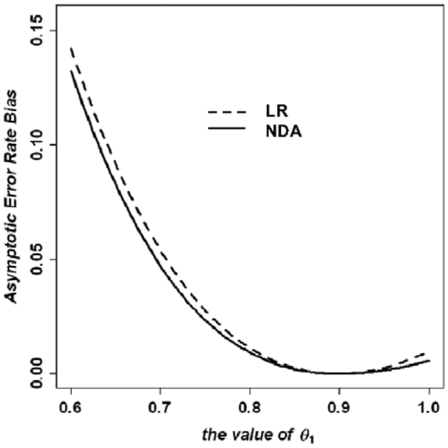

Fig. 2c. The asymptotic error rate bias with $\theta_0 = 0.9$, $\Delta = 3$ and $\pi_0 = 0.5$.

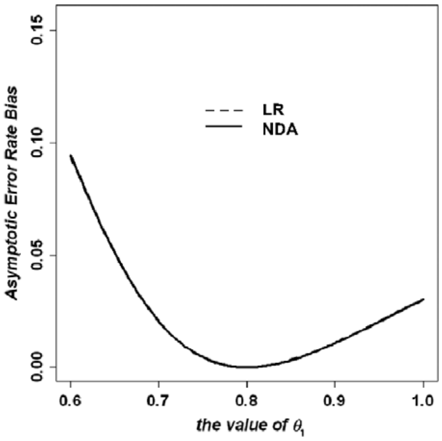

Fig. 3a. The asymptotic error rate bias with $\theta_0 = 0.8$, $\Delta = 1$ and $\pi_0 = 0.5$.

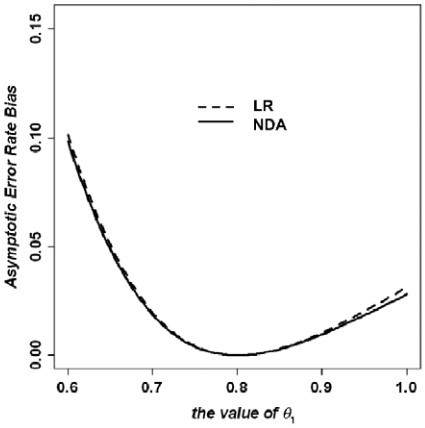

Fig. 3b. The asymptotic error rate bias with $\theta_0 = 0.8$, $\Delta = 2$ and $\pi_0 = 0.5$.

5. Summary

We have investigated the impact of CCC-Noise on the performance of a popular generative classifier, normal discriminant analysis (NDA) and its corresponding discriminative classifier logistic regression (LR). We compared the relative asymptotic error rate of these two classifiers under CCC-Noise when the underlying distributions are multivariate normal. Typically, the *AERB* of both procedures are only slightly affected when the noise level is low. LR has a larger *AERB* than NDA when the two populations are more separated.

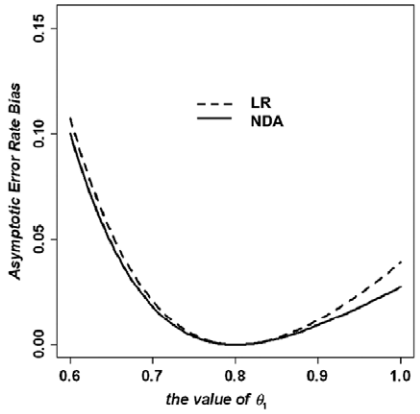

Fig. 3c. The asymptotic error rate bias with $\theta_0 = 0.8$, $\Delta = 3$ and $\pi_0 = 0.5$.

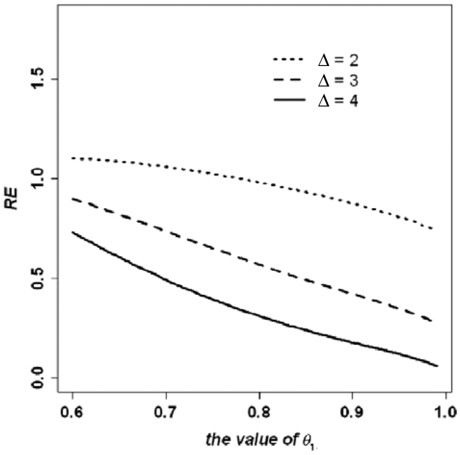

Fig. 4. The relative efficiency of LR to NDA with $\theta_0 = 1$, $n = 50$ and $\pi_0 = 0.5$.

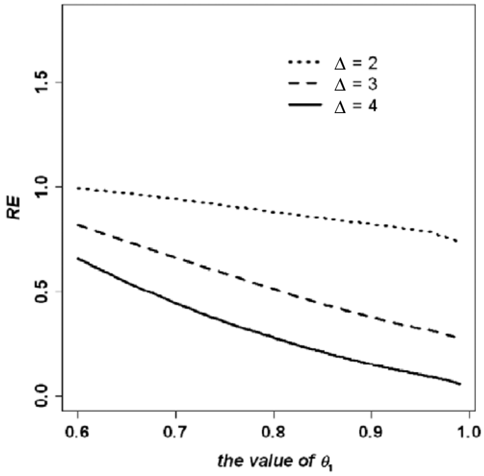

Fig. 5. The relative efficiency of LR to NDA with $\theta_0 = 1$, $n = 200$ and $\pi_0 = 0.5$.

With respect to the relative efficiency of LR to NDA, we showed that LR is less deteriorated by CCC-Noise compared to NDA. One important conclusion is that under the CCC-Noise contexts, the Mahalanobis distance Δ^2 plays a vital role in determining the relative performance of these two procedures. When Δ^2 is small, LR tends to be more tolerable to CCC-Noise compared to NDA.

Our analyses provide insight and a more in-depth understanding of the effect of noisy labeled observations on the accuracy of frequently used classification models, and conceivably our results can be used to guide interested researchers

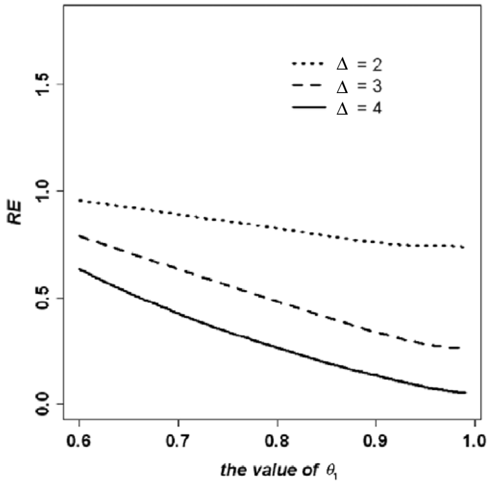

Fig. 6. The relative efficiency of LR to NDA with $\theta_0 = 1$, $n = 1000$ and $\pi_0 = 0.5$.

on the design of noise handling mechanisms. In future work we will extend this efficiency comparison to a non-Gaussian assumption for the distribution of **X**.

Appendix A. Proof of Lemma 2

According to White's theorem the misspecified MLEs, $(\hat{\lambda}, \hat{\boldsymbol{\mu}}_{\mathbf{0}}, \hat{\boldsymbol{\mu}}_{\mathbf{1}}, \hat{\boldsymbol{\Sigma}})$, of NDA converges to $(\lambda^*, \boldsymbol{\mu}_{\mathbf{0}}^*, \boldsymbol{\mu}_{\mathbf{1}}^*, \boldsymbol{\Sigma}^*)$ which minimizes

$$E_{(\mathbf{X},\tilde{Y})}\left\{(1-\tilde{Y})\left[\frac{p}{2}\log(2\pi)+\frac{1}{2}\log|\boldsymbol{\Sigma}|+\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{0}})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{0}})-\log\pi_0\right]\right.$$
$$\left.+\tilde{Y}\left[\frac{p}{2}\log(2\pi)+\frac{1}{2}\log|\boldsymbol{\Sigma}|+\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{1}})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{1}})-\log\pi_1\right]\right\},$$

where the expectation is with respect to the joint distribution of $(\tilde{Y}, \mathbf{X})$. Let L represent the inner term of the expectation. Taking the first partial derivatives with respect to λ, $\boldsymbol{\mu}_{\mathbf{0}}$, $\boldsymbol{\mu}_{\mathbf{1}}$ and $\boldsymbol{\Sigma}$ respectively and setting them to zero gives:

$$E_{(\mathbf{X},\tilde{Y})}\left\{\frac{\exp(\lambda^*)}{1+\exp(\lambda^*)}-\tilde{Y}\right\}=0, \tag{16}$$

$$E_{(\mathbf{X},\tilde{Y})}\left\{-(1-\tilde{Y})(\boldsymbol{\Sigma}^*)^{-1}(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{0}}^*)\right\}=\mathbf{0}, \tag{17}$$

$$E_{(\mathbf{X},\tilde{Y})}\left\{-\tilde{Y}(\boldsymbol{\Sigma}^*)^{-1}(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{1}}^*)\right\}=\mathbf{0}, \tag{18}$$

$$E_{(\mathbf{X},\tilde{Y})}\left\{-\frac{1}{2}[2\boldsymbol{\Sigma}^*-\mathrm{diag}\boldsymbol{\Sigma}^*]+\frac{1}{2}(1-\tilde{Y})[2(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{0}}^*)(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{0}}^*)'-\mathrm{diag}(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{0}}^*)(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{0}}^*)']\right.$$
$$\left.+\frac{1}{2}\tilde{Y}[2(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{1}}^*)(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{1}}^*)'-\mathrm{diag}(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{1}}^*)(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{1}}^*)']\right\}=\mathbf{0}. \tag{19}$$

Also it can be easily shown that, in the standard situation,

$$E_{(\mathbf{X},\tilde{Y})}[\mathbf{X}]=\frac{(\pi_1-\pi_0)\Delta\mathbf{e_1}}{2},\qquad E_{(\mathbf{X},\tilde{Y})}\left[\tilde{Y}\right]=\pi_0(1-\theta_0)+\pi_1\theta_1,\qquad E_{(\mathbf{X},\tilde{Y})}\left[\tilde{Y}\mathbf{X}\right]=\frac{[\theta_1\pi_1-(1-\theta_0)\pi_0]\,\Delta\mathbf{e_1}}{2},$$

$$E_{(\mathbf{X},\tilde{Y})}\left[(1-\tilde{Y})(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{0}}^*)(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{0}}^*)'\right]$$
$$=\pi_0\theta_0\left[\mathbf{I}+\left(\frac{\Delta\mathbf{e_1}}{2}+\boldsymbol{\mu}_{\mathbf{0}}^*\right)\left(\frac{\Delta\mathbf{e_1}}{2}+\boldsymbol{\mu}_{\mathbf{0}}^*\right)'\right]+\pi_1(1-\theta_1)\left[\mathbf{I}+\left(\frac{\Delta\mathbf{e_1}}{2}-\boldsymbol{\mu}_{\mathbf{0}}^*\right)\left(\frac{\Delta\mathbf{e_1}}{2}-\boldsymbol{\mu}_{\mathbf{0}}^*\right)'\right], \tag{20}$$

$$E_{(\mathbf{X},\tilde{Y})}\left[\tilde{Y}(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{1}}^*)(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{1}}^*)'\right]$$
$$=\pi_0(1-\theta_0)\left[\mathbf{I}+\left(\frac{\Delta\mathbf{e_1}}{2}+\boldsymbol{\mu}_{\mathbf{1}}^*\right)\left(\frac{\Delta\mathbf{e_1}}{2}+\boldsymbol{\mu}_{\mathbf{1}}^*\right)'\right]+\pi_1\theta_1\left[\mathbf{I}+\left(\frac{\Delta\mathbf{e_1}}{2}-\boldsymbol{\mu}_{\mathbf{1}}^*\right)\left(\frac{\Delta\mathbf{e_1}}{2}-\boldsymbol{\mu}_{\mathbf{1}}^*\right)'\right].$$

Substituting the values of the expressions from (20) into (16), (17), (18) and (19) and simplifying and solving yield $(\lambda^*, \boldsymbol{\mu}_\mathbf{0}^*, \boldsymbol{\mu}_\mathbf{1}^*, \boldsymbol{\Sigma}^*)$. Here $\boldsymbol{\Sigma}^* = \mathbf{I} + \left(\frac{ab}{a+b} + \frac{cd}{c+d}\right) \Delta^2 \mathbf{E_{11}}$ which indicates that $\boldsymbol{\sigma}^{(1)*} = h\mathbf{e_1}$.

The asymptotic variance–covariance matrix of $(\hat{\lambda}, \hat{\boldsymbol{\mu}}_\mathbf{0}, \hat{\boldsymbol{\mu}}_\mathbf{1}, \hat{\boldsymbol{\sigma}}^{(1)})$ is represented as

$$\boldsymbol{\Omega} = \mathbf{A^{-1}BA^{-1}} = \begin{pmatrix} A_{00} & \mathbf{A_{01}} & \mathbf{A_{02}} & \mathbf{A_{03}} \\ \mathbf{A}'_\mathbf{01} & \mathbf{A_{11}} & \mathbf{A_{12}} & \mathbf{A_{13}} \\ \mathbf{A}'_\mathbf{02} & \mathbf{A}'_\mathbf{12} & \mathbf{A_{22}} & \mathbf{A_{23}} \\ \mathbf{A}'_\mathbf{03} & \mathbf{A}'_\mathbf{13} & \mathbf{A}'_\mathbf{23} & \mathbf{A_{33}} \end{pmatrix}^{-1} \begin{pmatrix} B_{00} & \mathbf{B_{01}} & \mathbf{B_{02}} & \mathbf{B_{03}} \\ \mathbf{B}'_\mathbf{01} & \mathbf{B_{11}} & \mathbf{B_{12}} & \mathbf{B_{13}} \\ \mathbf{B}'_\mathbf{02} & \mathbf{B}'_\mathbf{12} & \mathbf{B_{22}} & \mathbf{B_{23}} \\ \mathbf{B}'_\mathbf{03} & \mathbf{B}'_\mathbf{13} & \mathbf{B}'_\mathbf{23} & \mathbf{B_{33}} \end{pmatrix} \mathbf{A^{-1}},$$

where A_{00} is a scalar, $\mathbf{A_{01}}, \mathbf{A_{02}}, \mathbf{A_{03}}$ are $1 \times p$ vectors and $\mathbf{A_{11}}, \mathbf{A_{22}}, \mathbf{A_{33}}, \mathbf{A_{12}}, \mathbf{A_{13}}, \mathbf{A_{23}}$ are $p \times p$ matrixes, likewise for all components in matrix **B**. First we consider the elements in matrix **A**. Again according to White's theorem we know that

$$A_{00} = E_{(\mathbf{X},\tilde{Y})}\left[\left.\frac{\partial^2 L}{\partial\lambda\partial\lambda}\right|_{(\lambda^*,\boldsymbol{\mu}_\mathbf{0}^*,\boldsymbol{\mu}_\mathbf{1}^*,\boldsymbol{\Sigma}^*)}\right] = \frac{\exp(\lambda^*)}{[1+\exp(\lambda^*)]^2},$$

$$\begin{aligned} \mathbf{A_{11}} &= E_{(\mathbf{X},\tilde{Y})}\left[\left.\frac{\partial^2 L}{\partial\boldsymbol{\mu}_\mathbf{0}\partial\boldsymbol{\mu}_\mathbf{0}}\right|_{(\lambda^*,\boldsymbol{\mu}_\mathbf{0}^*,\boldsymbol{\mu}_\mathbf{1}^*,\boldsymbol{\Sigma}^*)}\right] = E_{(\mathbf{X},\tilde{Y})}\left\{(1-Y')(\boldsymbol{\Sigma}^*)^{-1}\right\} \\ &= (a+b)\left(\mathbf{I} + \left(\frac{ab}{a+b} + \frac{cd}{c+d}\right)\Delta^2\mathbf{E_{11}}\right)^{-1}, \end{aligned}$$

$$\mathbf{A_{22}} = E_{(\mathbf{X},\tilde{Y})}\left[\left.\frac{\partial^2 L}{\partial\boldsymbol{\mu}_\mathbf{1}\partial\boldsymbol{\mu}_\mathbf{1}}\right|_{(\lambda^*,\boldsymbol{\mu}_\mathbf{0}^*,\boldsymbol{\mu}_\mathbf{1}^*,\boldsymbol{\Sigma}^*)}\right] = E_{(\mathbf{X},Y')}\left\{\tilde{Y}(\boldsymbol{\Sigma}^*)^{-1}\right\} = (c+d)\left(\mathbf{I} + \left(\frac{ab}{a+b} + \frac{cd}{c+d}\right)\Delta^2\mathbf{E_{11}}\right)^{-1}$$

and apparently $\mathbf{A_{01}} = \mathbf{A_{02}} = \mathbf{A_{03}} = \mathbf{O_{1\times p}}, \mathbf{A_{12}} = \mathbf{A_{21}} = \mathbf{O_{p\times p}}$. With respect to $\mathbf{A_{13}}$ and $\mathbf{A_{23}}$, for each element in $\sigma^{(1)}$, we have

$$E_{(\mathbf{X},\tilde{Y})}\left[\left.\frac{\partial^2 L}{\partial\boldsymbol{\mu}_\mathbf{0}\partial\sigma^{1j}}\right|_{(\lambda^*,\boldsymbol{\mu}_\mathbf{0}^*,\boldsymbol{\mu}_\mathbf{1}^*,\boldsymbol{\Sigma}^*)}\right] = E_{(\mathbf{X},\tilde{Y})}\left[\frac{\mathbf{E_{1j}}+\mathbf{E_{j1}}}{1+\delta_{1j}}(1-\tilde{Y})(\mathbf{X}-\boldsymbol{\mu}_\mathbf{0}^*)\right] = \mathbf{O_{p\times 1}},$$

$$E_{(\mathbf{X},\tilde{Y})}\left[\left.\frac{\partial^2 L}{\partial\boldsymbol{\mu}_\mathbf{1}\partial\boldsymbol{\sigma}^\mathbf{1j}}\right|_{(\lambda^*,\boldsymbol{\mu}_\mathbf{0}^*,\boldsymbol{\mu}_\mathbf{1}^*,\boldsymbol{\Sigma}^*)}\right] = E_{(\mathbf{X},\tilde{Y})}\left[\frac{\mathbf{E_{1j}}+\mathbf{E_{j1}}}{1+\delta_{1j}}\tilde{Y}(\mathbf{X}-\boldsymbol{\mu}_\mathbf{1}^*)\right] = \mathbf{O_{p\times 1}},$$

which indicate that $\mathbf{A_{13}} = \mathbf{A_{23}} = \mathbf{O_{p\times p}}$.

Now let's consider (19). Taking the first partial derivative with respect to σ^{ij},and evaluated at $(\lambda^*, \boldsymbol{\mu}_\mathbf{0}^*, \boldsymbol{\mu}_\mathbf{1}^*, \boldsymbol{\Sigma}^*)$ gives

$$-E_{(\mathbf{X},\tilde{Y})}\left[-\boldsymbol{\Sigma}^*\frac{\mathbf{E_{ij}}+\mathbf{E_{ji}}}{1+\delta_{ij}}\boldsymbol{\Sigma}^* + \frac{1}{2}\mathrm{diag}\left(\boldsymbol{\Sigma}^*\frac{\mathbf{E_{ij}}+\mathbf{E_{ji}}}{1+\delta_{ij}}\boldsymbol{\Sigma}^*\right)\right].$$

All the elements in this matrix are zero except the positions (i, j) and (j, i), which indicates that $\mathbf{A_{33}}$ is a diagonal matrix and evaluation gives $\mathbf{A_{33}} = \frac{h^2}{2}\mathbf{E_{11}} + h(\mathbf{I} - \mathbf{E_{11}})$. Finally, the whole matrix **A** can be written in the following way

$$\begin{aligned} \mathbf{A} = \mathrm{diag}\Bigg(& -\frac{\exp(\lambda^*)}{[1+\exp(\lambda^*)]^2}, -\frac{a+b}{h}, \overbrace{-(a+b), \ldots, -(a+b)}^{p-1}, \\ & -\frac{c+d}{h}, \overbrace{-(c+d), \ldots, -(c+d)}^{p-1}, \frac{-h^2}{2}, \overbrace{-h/2, \ldots, -h/22}^{p-1}\Bigg). \end{aligned}$$

Now consider the matrix **B**. According to White's theorem we know that

$$B_{00} = E_{(\mathbf{X},\tilde{Y})}\left[\left.\frac{\partial L}{\partial\lambda}\frac{\partial L}{\partial\lambda}\right|_{(\lambda^*,\boldsymbol{\mu}_\mathbf{0}^*,\boldsymbol{\mu}_\mathbf{1}^*,\boldsymbol{\Sigma}^*)}\right] = E_{(\mathbf{X},\tilde{Y})}\left\{\left[-\frac{\exp(\lambda)}{1+\exp(\lambda)} + \tilde{Y}\right]^2\right\} = \frac{\exp(\lambda^*)}{[1+\exp(\lambda^*)]^2},$$

$$\begin{aligned} \mathbf{B_{11}} &= E_{(\mathbf{X},\tilde{Y})}\left[\left.\frac{\partial L}{\partial\boldsymbol{\mu}_\mathbf{0}}\left(\frac{\partial L}{\partial\boldsymbol{\mu}_\mathbf{0}}\right)'\right|_{(\lambda^*,\boldsymbol{\mu}_\mathbf{0}^*,\boldsymbol{\mu}_\mathbf{1}^*,\boldsymbol{\Sigma}^*)}\right] = E_{(\mathbf{X},Y')}\left\{(1-\tilde{Y})^2(\boldsymbol{\Sigma}^*)^{-1}(\mathbf{X}-\boldsymbol{\mu}_\mathbf{0}^*)(\mathbf{X}-\boldsymbol{\mu}_\mathbf{0}^*)'(\boldsymbol{\Sigma}^*)^{-1}\right\} \\ &= \left(\mathbf{I} + \left(\frac{ab}{a+b} + \frac{cd}{c+d}\right)\Delta^2\mathbf{E_{11}}\right)^{-1}\left[(a+b)\mathbf{I} + \frac{ab}{a+b}\Delta^2\mathbf{E_{11}}\right]\left(\mathbf{I} + \left(\frac{ab}{a+b} + \frac{cd}{c+d}\right)\Delta^2\mathbf{E_{11}}\right)^{-1}, \end{aligned}$$

$$\begin{aligned} \mathbf{B_{22}} &= E_{(\mathbf{X},\tilde{Y})}\left[\left.\frac{\partial L}{\partial\boldsymbol{\mu}_\mathbf{1}}\left(\frac{\partial L}{\partial\boldsymbol{\mu}_\mathbf{1}}\right)'\right|_{(\lambda^*,\boldsymbol{\mu}_\mathbf{0}^*,\boldsymbol{\mu}_\mathbf{1}^*,\boldsymbol{\Sigma}^*)}\right] = E_{(\mathbf{X},Y')}\left\{(Y')^2(\boldsymbol{\Sigma}^*)^{-1}(\mathbf{X}-\boldsymbol{\mu}_\mathbf{1}^*)(\mathbf{X}-\boldsymbol{\mu}_\mathbf{1}^*)'(\boldsymbol{\Sigma}^*)^{-1}\right\} \\ &= \left(\mathbf{I} + \left(\frac{ab}{a+b} + \frac{cd}{c+d}\right)\Delta^2\mathbf{E_{11}}\right)^{-1}\left[(c+d)\mathbf{I} + \frac{cd}{c+d}\Delta^2\mathbf{E_{11}}\right]\left(\mathbf{I} + \left(\frac{ab}{a+b} + \frac{cd}{c+d}\right)\Delta^2\mathbf{E_{11}}\right)^{-1}. \end{aligned}$$

It also can be easily shown that $\mathbf{B_{01}} = \mathbf{B_{02}} = \mathbf{O_{1\times p}}$ and $\mathbf{B_{12}} = \mathbf{B_{21}} = \mathbf{O_{p\times p}}$ based on the fact that $(1-\tilde{Y})\tilde{Y}$ is a zero random variable. Now let's consider $\mathbf{B_{03}}$. First we let $\mathbf{N} = -\frac{1}{2}\left[2\boldsymbol{\Sigma}^* - \operatorname{diag}\boldsymbol{\Sigma}^*\right]$ and n_{ij} be (i,j) component of $\mathbf{N}$. Thus we have

$$E_{(\mathbf{X},\tilde{Y})}\left[\frac{\partial L}{\sigma^{ij}}\bigg|_{(\lambda^*,\boldsymbol{\mu}_\mathbf{0}^*,\boldsymbol{\mu}_\mathbf{1}^*,\boldsymbol{\Sigma}^*)}\right] = n_{ij} + (1-\tilde{Y})\frac{(X_i-\mu_{0i}^*)(X_j-\mu_{0j}^*)}{1+\delta_{ij}} + \tilde{Y}\frac{(X_i-\mu_{1i}^*)(X_j-\mu_{1j}^*)}{1+\delta_{ij}}.$$

Note the fact that $\mathbf{N}$ is a diagonal matrix. Let's consider the jth component of $\mathbf{B_{03}}$

$$\begin{aligned}E_{(\mathbf{X},\tilde{Y})}\left[\frac{\partial L}{\lambda}\frac{\partial L}{\sigma^{1j}}\bigg|_{(\lambda^*,\boldsymbol{\mu}_\mathbf{0}^*,\boldsymbol{\mu}_\mathbf{1}^*,\boldsymbol{\Sigma}^*)}\right] &= E_{(\mathbf{X},\tilde{Y})}\left\{\left(\frac{\exp(\lambda)}{1+\exp(\lambda)} - \tilde{Y}\right)\right.\\ &\quad\left.\times\left[n_{1j} + (1-\tilde{Y})\frac{(X_1-\mu_{01}^*)(X_j-\mu_{0j}^*)}{1+\delta_{1j}} + \tilde{Y}\frac{(X_1-\mu_{11}^*)(X_j-\mu_{1j}^*)}{1+\delta_{1j}}\right]\right\}\\ &= E_{(\mathbf{X},\tilde{Y})}\left\{-\tilde{Y}n_{1j} - \tilde{Y}\frac{(X_1-\mu_{11}^*)(X_j-\mu_{1j}^*)}{1+\delta_{1j}}\right\}.\end{aligned}$$

It can be observed that only the component with $j=1$ is nonzero, which is equal to

$$E_{(\mathbf{X},\tilde{Y})}\left\{\tilde{Y}\frac{h}{2} - \tilde{Y}\frac{(X_1-\mu_{11}^*)(X_1-\mu_{11}^*)}{2}\right\} = \frac{\Delta^2}{2}\left[ab\frac{c+d}{a+b} - cd\frac{a+b}{c+d}\right].$$

Now let's consider the (i,j) component of $\mathbf{B_{33}}$,

$$\begin{aligned}E_{(\mathbf{X},\tilde{Y})}\left[\frac{\partial L}{\sigma^{1j}}\frac{\partial L}{\sigma^{1i}}\bigg|_{(\lambda^*,\boldsymbol{\mu}_\mathbf{0}^*,\boldsymbol{\mu}_\mathbf{1}^*,\boldsymbol{\Sigma}^*)}\right] &= E\left[(1-\tilde{Y})^2\frac{(X_1-\mu_{01}^*)(X_j-\mu_{0j}^*)}{1+\delta_{1j}}\frac{(X_1-\mu_{01}^*)(X_i-\mu_{0i}^*)}{1+\delta_{1i}}\right]\\ &\quad+E\left[(\tilde{Y})^2\frac{(X_1-\mu_{11}^*)(X_j-\mu_{1j}^*)}{1+\delta_{1j}}\frac{(X_1-\mu_{11}^*)(X_i-\mu_{1i}^*)}{1+\delta_{1i}}\right].\end{aligned}$$

It can be observed that only the components with i equal to j are nonzero. For any $j \neq 1$

$$\begin{aligned}E_{(\mathbf{X},\tilde{Y})}\left[\frac{\partial L}{\sigma^{1j}}\frac{\partial L}{\sigma^{1j}}\bigg|_{(\lambda^*,\boldsymbol{\mu}_\mathbf{0}^*,\boldsymbol{\mu}_\mathbf{1}^*,\boldsymbol{\Sigma}^*)}\right] &= -n_{1j}^2 + 1 + \pi_0\theta_0(\mu_{01}-\mu_{01}^*)^2 + \pi_1(1-\theta_1)(\mu_{11}-\mu_{01}^*)^2\\ &\quad+\pi_0(1-\theta_0)(\mu_{01}-\mu_{11}^*)^2 + \pi_1\theta_1(\mu_{11}-\mu_{11}^*)^2\\ &= -n_{1j}^2 + 1 + \left(\frac{\Delta}{2}\right)^2\left(a\left(\frac{2b}{b+a}\right)^2 + b\left(\frac{2a}{b+a}\right)^2\right.\\ &\quad\left.+c\left(\frac{2d}{c+d}\right)^2 + d\left(\frac{2c}{c+d}\right)^2\right)\\ &= 1 + \Delta^2\left[\frac{ab}{b+a} + \frac{cd}{c+d}\right] = h.\end{aligned}$$

For $j=1$,

$$E_{(\mathbf{X},\tilde{Y})}\left[\frac{\partial L}{\sigma^{11}}\frac{\partial L}{\sigma^{11}}\bigg|_{(\lambda^*,\boldsymbol{\mu}_\mathbf{0}^*,\boldsymbol{\mu}_\mathbf{1}^*,\boldsymbol{\Sigma}^*)}\right] = \frac{h^2}{2} + \frac{1}{4}\Delta^4\left[\frac{ab^4+a^4b}{(b+a)^4} + \frac{cd^4+c^4d}{(d+c)^4} - 3\left(\frac{ab}{b+a} + \frac{cd}{d+c}\right)^2\right].$$

For (i,j) component of $\mathbf{B_{13}}$,

$$E_{(\mathbf{X},\tilde{Y})}\left[\frac{\partial L}{\sigma^{1j}}\frac{\partial L}{\partial\mu_{0i}}\bigg|_{(\lambda^*,\boldsymbol{\mu}_\mathbf{0}^*,\boldsymbol{\mu}_\mathbf{1}^*,\boldsymbol{\Sigma}^*)}\right] = E_{(\mathbf{X},\tilde{Y})}\left[(1-\tilde{Y})^2\frac{(X_1-\mu_{01}^*)(X_j-\mu_{0j}^*)}{1+\delta_{1j}}\frac{(X_i-\mu_{0i}^*)}{h^{\delta_{1i}}}\right].$$

It can be observed that only the $(1,1)$ component is nonzero, which is equal to

$$\begin{aligned}E_{(\mathbf{X},\tilde{Y})}\left[\frac{\partial L}{\sigma^{11}}\frac{\partial L}{\partial\mu_{01}}\bigg|_{(\lambda^*,\boldsymbol{\mu}_\mathbf{0}^*,\boldsymbol{\mu}_\mathbf{1}^*,\boldsymbol{\Sigma}^*)}\right] &= \pi_0\theta_0\frac{3(\mu_{01}-\mu_{01}^*) + (\mu_{01}-\mu_{01}^*)^3}{2h}\\ &\quad+\pi_1(1-\theta_1)\frac{3(\mu_{11}-\mu_{01}^*) + (\mu_{11}-\mu_{01}^*)^3}{2h}\\ &= \frac{\Delta^3}{2h}\frac{ab(a-b)}{(a+b)^2}.\end{aligned}$$

Thus $\mathbf{B_{13}} = \frac{\Delta^3}{2h}\frac{ab(a-b)}{(a+b)^2}\mathbf{E_{11}}$. Similarly, we have $\mathbf{B_{23}} = \frac{\Delta^3}{2h}\frac{cd(c-d)}{(c+d)^2}\mathbf{E_{11}}$. Finally, evaluation of $\mathbf{A^{-1}BA^{-1}}$ gives the matrix $\boldsymbol{\Omega}$.

Appendix B. Proof of Lemma 4

The asymptotic variance–covariance matrix of $(\hat{s}, \hat{\mathbf{t}}')$ is represented

$$\mathbf{\Psi} = \mathbf{V^{-1}HV^{-1}} = \begin{pmatrix} V_{11} & \mathbf{V_{12}} \\ \mathbf{V}'_{\mathbf{12}} & \mathbf{V_{22}} \end{pmatrix}^{-1} \begin{pmatrix} H_{11} & \mathbf{H_{12}} \\ \mathbf{H}'_{\mathbf{12}} & \mathbf{H_{22}} \end{pmatrix} \mathbf{V^{-1}}$$

where V_{11} is a scalar, $\mathbf{V_{12}}$ are $1 \times p$ vectors and $\mathbf{V_{22}}$ are $p \times p$ matrixes, likewise for all components in matrix $\mathbf{H}$. First we consider the elements in matrix $\mathbf{H}$. Let l represent the inner term of the expectation of (10),

$$H_{11} = E_{(\mathbf{X},\tilde{Y})}\left[\left.\frac{\partial^2 l}{\partial\alpha\partial\alpha}\right|_{(s^*,u^*)}\right] = E_{X_1}\left\{\frac{\exp(s^*)\exp(u^*\Delta X_1)}{[1+\exp(s^*)\exp(u^*\Delta X_1)]^2}\right\} = h_0.$$

Now consider the ith component of $\mathbf{H_{12}}$

$$E_{(\mathbf{X},\tilde{Y})}\left[\left.\frac{\partial^2 l}{\partial\alpha\partial\beta_i}\right|_{(s^*,u^*)}\right] = E_{\mathbf{X}}\left\{\frac{X_i\exp(s^*)\exp(u^*\Delta X_1)}{[1+\exp(s^*)\exp(u^*\Delta X_1)]^2}\right\}.$$

Only the first component is nonzero, which is equal to

$$E_{(\mathbf{X},\tilde{Y})}\left[\left.\frac{\partial^2 l}{\partial\alpha\partial\beta_1}\right|_{(s^*,u^*)}\right] = E_{X_1}\left\{\frac{X_1\exp(s^*)\exp(u^*\Delta X_1)}{[1+\exp(s^*)\exp(u^*\Delta X_1)]^2}\right\} = h_1.$$

The (i, j) component of $\mathbf{H_{22}}$,

$$E_{(\mathbf{X},\tilde{Y})}\left[\left.\frac{\partial^2 l}{\partial\beta_i\partial\beta_j}\right|_{(s^*,u^*)}\right] = E_{\mathbf{X}}\left\{\frac{X_iX_j\exp(s^*)\exp(u^*\Delta X_1)}{[1+\exp(s^*)\exp(u^*\Delta X_1)]^2}\right\}.$$

It can be observed that only the components with i equal to j are nonzero. For any $i \neq 1$

$$\begin{aligned} E_{(\mathbf{X},\tilde{Y})}\left[\left.\frac{\partial^2 l}{\partial\beta_i\partial\beta_i}\right|_{(s^*,u^*)}\right] &= E_{\mathbf{X}}\left\{\frac{(X_i)^2\exp(s^*)\exp(u^*\Delta X_1)}{[1+\exp(s^*)\exp(u^*\Delta X_1)]^2}\right\} \\ &= E_{X_1}\left\{\frac{\exp(s^*)\exp(u^*\Delta X_1)}{[1+\exp(s^*)\exp(u^*\Delta X_1)]^2}\right\} = h_0. \end{aligned}$$

For $i = 1$

$$E_{(\mathbf{X},\tilde{Y})}\left[\left.\frac{\partial^2 l}{\partial\beta_1\partial\beta_1}\right|_{(s^*,u^*)}\right] = E_{\mathbf{X}}\left\{\frac{(X_1)^2\exp(s^*)\exp(u^*\Delta X_1)}{[1+\exp(s^*)\exp(u^*\Delta X_1)]^2}\right\} = h_2.$$

Thus $\mathbf{H_{22}} = h_2\mathbf{E_{11}} + h_0(\mathbf{I} + \mathbf{E_{11}})$. For matrix $\mathbf{V}$,

$$\begin{aligned} V_{11} &= E_{(\mathbf{X},\tilde{Y})}\left[\left.\frac{\partial l}{\partial\alpha}\frac{\partial l}{\partial\alpha}\right|_{(s^*,u^*)}\right] \\ &= E_{(\mathbf{X},\tilde{Y})}\left\{(Y')^2 - 2\frac{\exp(s^*)\exp(u^*\Delta X_1)}{1+\exp(s^*)\exp(u^*\Delta X_1)}Y' + \left[\frac{\exp(s^*)\exp(u^*\Delta X_1)}{1+\exp(s^*)\exp(u^*\Delta X_1)}\right]^2\right\} \\ &= E_{X_1}\left[\frac{\theta_1\pi_1\exp(\Delta X_1) + \pi_0(1-\theta_0)}{\pi_0 + \pi_1\exp(\Delta X_1)}\frac{1-\exp(s^*)\exp(u^*\Delta X_1)}{1+\exp(s^*)\exp(u^*\Delta X_1)} + \left[\frac{\exp(s^*)\exp(u^*\Delta X_1)}{1+\exp(s^*)\exp(u^*\Delta X_1)}\right]^2\right] = v_0. \end{aligned}$$

For the ith component of $\mathbf{V_{12}}$,

$$E_{(\mathbf{X},\tilde{Y})}\left[\left.\frac{\partial l}{\partial\alpha}\frac{\partial l}{\partial\beta_i}\right|_{(s^*,u^*)}\right] = E_{(\mathbf{X},\tilde{Y})}\left\{(Y')^2X_i - 2\frac{\exp(s^*)\exp(u^*\Delta X_1)}{1+\exp(s^*)\exp(u^*\Delta X_1)}Y'X_i + X_i\left[\frac{\exp(s^*)\exp(u^*\Delta X_1)}{1+\exp(s^*)\exp(u^*\Delta X_1)}\right]^2\right\}.$$

Only the first component is nonzero, which is $E_{(\mathbf{X},\tilde{Y})}\left[\left.\frac{\partial l}{\partial\alpha}\frac{\partial l}{\partial\beta_1}\right|_{(s^*,u^*)}\right] = v_1$. For the matrix $\mathbf{V_{22}}$, similar to $\mathbf{H_{22}}$, we have $\mathbf{V_{22}} = v_2\mathbf{E_{11}} + v_0(I + \mathbf{E_{11}})$. Finally, evaluation of $\mathbf{V^{-1}HV^{-1}}$ gives the matrix $\mathbf{\Psi}$.

References

[1] T. Krishnan, S.C. Nandy, Efficiency of discriminant analysis when initial samples are classified stochastically, Pattern Recognition 23 (1990) 529–537.
[2] U.A. Katre, T. Krishnan, Pattern recognition with imperfect supervision, Pattern Recognition 22 (1989) 423–431.
[3] J.E. Michalek, R.C. Tripathi, The effect of errors in diagnosis and measurement on the estimation of the probability of an event, Journal of the American Statistical Association 75 (1980) 713–721.
[4] Y. Li, W. Lodeswyk, D. De Ridder, M. Reinders, Classification in the presence of class noise using a probabilistic kernel fisher method, Pattern Recognition 40 (2007) 3349–3357.
[5] N.D. Lawrence, B. Scholkopf, Estimating a kernel fisher discriminant in the presence of label noise, in: Proc. of 18th ICML, 2001, pp. 306–313.
[6] Y. Yasui, M. Pepe, L. Hsu, B.L. Adam, Z. Feng, Partially supervised learning using an EM-boosting algorithm, Biometrics 60 (2004) 199–206.
[7] W. Zhang, R. Rekaya, K. Bertrand, A method for predicting disease subtypes in presence of misclassification among training samples using gene expression: application to human breast cancer, Bioinformatics 22 (2006) 317–325.
[8] K. Robbins, S. Joseph, W. Zhang, R. Rekaya, J.K. Bertrand, Classification of incipient Alzheimer patients using gene expression data: dealing with potential misdiagnosis, Online Journal of Bioinformatics 7 (2006) 22–31.
[9] J. Maletic, A. Marcus, Data cleansing: beyond integrity analysis, in: Proceedings of the Conference on Information Quality, 2000.
[10] S. Weisberg, Applied Linear Regression, John Wiley and Sons, Inc., 1980.
[11] S.W. Norton, H. Hirsh, Classifier learning from noisy data as probabilistic evidence combination, in: Proceedings of the Tenth National Conference on Artificial Intelligence, AAAI Press, Menlo Park, CA, 1992, pp. 141–146.
[12] L.S. Magder, J.P. Hughes, Logistic regression when the outcome is measured with uncertainty, American Journal of Epidemiology 146 (1997) 195–203.
[13] M. Hofler, The effect of misclassification on the estimation of association: a review, International Journal of Methods in Psychiatric Research 14 (2005) 92–101.
[14] J.M. Neuhaus, Bias and efficiency loss due to misclassified responses in binary regression, Biometrika 86 (1999) 843–855.
[15] T. Krishnan, Efficiency of learning with imperfect supervision, Pattern Recognition 21 (1988) 183–188.
[16] B. Efron, The efficiency of logistic regression compared to normal discriminant analysis, Journal of the American Statistical Association 70 (1975) 892–898.
[17] X. Zhu, X. Wu, Class noise vs. attribute noise: a quantitative study of their impacts, Artificial Intelligence Review 22 (2004) 177–210.
[18] A.Y. Ng, M.I. Jordan, On discriminative vs. generative classifiers: a comparison of logistic regression and naïve Bayes, in: NIPS, 2001, pp. 841–848.
[19] Y. Bi, Theoretical analysis of classification under CCC-Noise and its application to semi-supervised text classification, Ph.D. Thesis, University of California, Riverside, 2008.
[20] H. White, Maximum likelihood estimation of misspecified model, Econometrica 50 (1982) 1–25.
[21] K.C. Li, N. Duan, Regression analysis under link violation, Annals of Statistics 17 (1989) 1009–1052.