Binary Naive Bayesian classifiers for correlated Gaussian features: A theoretical analysis

Ewald van Dyk[1], *Etienne Barnard*[2]

[1,2]School of Electrical, Electronic and Computer Engineering, University of North-West, South Africa
[1,2]Human Language Technologies Research Group, Meraka Institute, Pretoria, South Africa
evdyk@csir.co.za, ebarnard@csir.co.za

Abstract

We investigate the use of Naive Bayesian classifiers for correlated Gaussian feature spaces and derive error estimates for these classifiers. The error analysis is done by developing an exact expression for the error performance of a binary classifier with Gaussian features while using any quadratic decision boundary. Therefore, the analysis is not restricted to Naive Bayesian classifiers alone and can, for instance, be used to calculate the Bayes error performance. We compare the analytical error rate to that obtained when Monte-Carlo simulations are performed for a 2 and 12 dimensional binary classification problem. Finally, we illustrate the robust performances obtained with Naive Bayesian classifiers (as apposed to a maximum likelihood classifier) for high dimensional problems when data sparsity becomes an issue.

1. Introduction

The popularity of Naive Bayesian (NB) classifiers has increased in recent years [1, 2], among others due to exceptional classification performance in high dimensional feature spaces. NB classifiers ignore all correlation between features and are inexpensive to use in high dimensional spaces where it becomes practically infeasible to estimate accurate correlation parameters. An attempt to estimate correlations can often lead to over fitting and decrease the performance (both efficiency and accuracy) of the classifier. Empirical evidence and an intuitive explanation on why NB classifiers perform so well in high dimensional feature spaces (in terms of the bias-variance problem) can be found in [3].

The increase in popularity of NB classifiers has not been matched by a similar growth in theoretical understanding (such as proper error analysis and feature selection). In one of our previous papers [2], we developed analytical tools for estimating error rates and used them as similarity measures for feature selection in discrete environments (all features were assumed to be multinomial).

In this paper, we focus on developing an exact expression for the error rates of binary (two-class) NB classifiers where all features are continuous, correlated multivariate Gaussian distributions.

There have been a few misunderstandings in the past regarding NB classifiers. One good example as pointed out by [3] is the confusion between NB classifiers and linear classifiers in [4]. Consider, for example, a parametric classifier where all features are assumed to be Gaussian. The only way that one can obtain a piecewise linear boundary, is if all classes have identical covariance matrices, which is clearly not the case for general NB classifiers. Therefore, later on in this paper, we discuss the different decision boundaries that can be obtained in a binary NB classification problem with Gaussian features and discuss their intuitive meaning.

In order to calculate the error performance of a binary NB classifier we turn to basic decision theory were we calculate an NB decision boundary that separates two hyperspace partitions Ω_1 and Ω_2. Whenever an observed feature vector falls within region Ω_1 or Ω_2, we classify the pattern to come from class ω_1 or ω_2 respectively. Therefore we can calculate the classification error rate by computing eq. 1[5]

$$\epsilon = p(\omega_1) \int_{\Omega_2} p(\mathbf{x}|\omega_1) d\mathbf{x} + p(\omega_2) \int_{\Omega_1} p(\mathbf{x}|\omega_2) d\mathbf{x}, \quad (1)$$

where ϵ is the classification error rate, $\mathbf{x}$ is the input vector and $p(\omega_1)$ and $p(\omega_2)$ are the prior probabilities for classes ω_1 and ω_2 respectively. Therefore, the very specific challenge addressed in this paper, is to calculate the integral parts in eq. 1, where $p(\mathbf{x}|\omega_1)$ and $p(\mathbf{x}|\omega_2)$ are correlated Gaussian distributions of arbitrary dimensionality. Since we are working with NB classifiers, the decision boundary will generally be a quadratic surface.

There exist many upper bounds on the Bayes error rate for Gaussian classification problems. Some popular loose bounds that can be calculated efficiently include the Chernoff bound [6] and the Bhattacharyya bound [7]. Some tighter upper bounds include the equivocation bound [8], Bayesian distance bound [9], sinusoidal bound [10] and exponential bound [11]. Unfortunately, none of these bounds are useful for the analysis of NB classifiers, since they obtain bounds for the Bayes error rate which do not allow us to investigate the effects of the assumption of uncorrelatedness. In order to investigate these effects, we choose to calculate an asymptotically exact error rate. The easiest way to do this, is to do Monte-Carlo simulations where we generate samples from the class distributions and simply count the errors; this is a time-consuming exercise, but does asymptotically converge to the true error rate. Instead, we derive an exact analytical expression similar to work done in [12, 13]. In our derivation, we first transform the integral problems in eq. 1 into a problem of finding the cumulative distribution (cdf) of a linear combination of chi-square variates.

The main contribution of the current paper is that we are able to derive exact analytic expressions for the Naive Bayesian error rate in the general case, whereas previous authors were able to do so only in terms of computationally expensive series expansions [14] or imprecise approximations [13].

The rest of this paper is organized as follows. In section 2, we derive the equations needed to transform the classification problem into one represented as a linear combination of chi-square variates. In section 3, we discuss all possible quadratic

decision boundaries obtained in the context of the work done in section 2 and we show the exact solution to the cdf for most of these boundaries. In section 4, we run simulations to compare NB error rates obtained from both Monte-Carlo simulations and the analytical expressions found.

None of the theory developed in sections 2 and 3 is limited to NB classifiers and applies to quadratic discriminant analysis (QDA) in general. To be more specific, Sections 2 and 3 focus on methods for calculating $\int_{\Omega_2} p(\mathbf{x}|\omega_1)d\mathbf{x}$. It is easy to calculate $\int_{\Omega_1} p(\mathbf{x}|\omega_2)d\mathbf{x}$ by simply reversing the roles of ω_1 and ω_2.

2. Linear combinations of non-central chi-square variates

Let us assume that $p(\mathbf{x}|\omega_1)$ and $p(\mathbf{x}|\omega_2)$ are both Gaussian distributions with means μ_1 and μ_2 and covariance matrices Σ_1 and Σ_2 respectively. Therefore

$$p(\mathbf{x}|\omega_i) = \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma_i}|^{1/2}} \exp(-\frac{1}{2}(\mathbf{x}-\mu_i)^T\mathbf{\Sigma_i^{-1}}(\mathbf{x}-\mu_i)), \tag{2}$$

where D is the dimensionality of the problem. Unfortunately, the exact values for μ_i and Σ_i are almost never known and need to be estimated, with say $\hat{\mu}_i$ and $\hat{\mathbf{\Sigma}}_i$. For NB classifiers, $\hat{\mathbf{\Sigma}}_i$ is a diagonal matrix. For simplicity we assume that $\hat{\mu}_1 = \mu_1$ and $\hat{\mu}_2 = \mu_2$ - inaccuracy in estimating the sample means is best treated as a separate issue.

We can use these estimates to calculate the decision boundary for a binary classification problem. Eq.3 is the simplest way to describe the decision boundary hyperplane in terms of the estimated parameters.

$$p(\omega_1)p(\mathbf{x}|\hat{\mu}_1, \hat{\mathbf{\Sigma}}_1) = p(\omega_2)p(\mathbf{x}|\hat{\mu}_2, \hat{\mathbf{\Sigma}}_2) \tag{3}$$

When we take the logarithm on both sides of eq. 3 and use eq. 2, we get the following representation for the decision boundary:

$$\beta(\mathbf{x}) = (\mathbf{x}-\hat{\mu}_1)^T\hat{\mathbf{\Sigma}}_1^{-1}(\mathbf{x}-\hat{\mu}_1) - (\mathbf{x}-\hat{\mu}_2)^T\hat{\mathbf{\Sigma}}_2^{-1}(\mathbf{x}-\hat{\mu}_2) = t_1, \tag{4}$$

where

$$t_1 = \log(\frac{|\hat{\mathbf{\Sigma}}_2|}{|\hat{\mathbf{\Sigma}}_1|}) + 2\log(\frac{p(\omega_1)}{p(\omega_1)}).$$

In the context of eq. 1, it is easy to see that

$$\int_{\Omega_2} p(\mathbf{x}|\omega_1)d\mathbf{x} = p(\beta_1(\mathbf{x}) \geq t_1), \tag{5}$$

where $\mathbf{x} \sim N(\mu_1, \mathbf{\Sigma_1})$.

In the rest of this section, we focus our efforts on transforming eq. 4 into a much more usable form,

$$F(\Phi, \mathbf{m}, t) = p(\sum_{i=1}^{D} \phi_i(y_i - m_i)^2 \leq t), \tag{6}$$

where $\mathbf{y} \sim N(\mathbf{0}, \mathbf{I})$, $F(\Phi, \mathbf{m}, t)$ is a function that we can relate to the error (see section 3), ϕ_i and m_i are variance and bias constants. We do the transformation in four steps as follows.

2.1. Shift means by μ_1

We define $\mathbf{z} = \mathbf{x} - \mu_1$ and with a little manipulation (and assuming $\hat{\mu}_1 = \mu_1$ and $\hat{\mu}_2 = \mu_2$) we can rewrite eq. 4 as follow.

$$\begin{aligned}
\beta_2(\mathbf{z}) &= \mathbf{z}^T\mathbf{B_1}\mathbf{z} - 2\mathbf{b_1^T}\mathbf{z} = t_2 \\
\mathbf{B_1} &= \hat{\mathbf{\Sigma}}_\mathbf{1}^{-1} - \hat{\mathbf{\Sigma}}_\mathbf{2}^{-1} \\
\mathbf{b_1^T} &= (\mu_1 - \mu_2)^T\hat{\mathbf{\Sigma}}_\mathbf{2}^{-1} \\
t_2 &= t_1 + (\mu_1 - \mu_2)^T\hat{\mathbf{\Sigma}}_\mathbf{2}^{-1}(\mu_1 - \mu_2) \\
\mathbf{z} &\sim N(\mathbf{0}, \mathbf{\Sigma_1})
\end{aligned} \tag{7}$$

Note that $\mathbf{B_1}$ is in general not a positive-definite matrix, but is symmetric and can be rotated.

2.2. Rotate matrices to Diagonalize $\mathbf{\Sigma_1}$

Since $\mathbf{z}$ is centered at the origin, we can rotate $\mathbf{\Sigma_1}$ to be diagonal, as long as we rotate the decision boundary too. We define $\mathbf{v} = \mathbf{U}_{\omega_1}^\mathbf{T}\mathbf{z}$, where $\mathbf{U}_{\omega_1}$ is the eigenvector matrix of $\mathbf{\Sigma_1}$ satisfying
$\mathbf{U}_{\omega_1}^\mathbf{T}\mathbf{\Sigma_1}\mathbf{U}_{\omega_1} = \mathbf{\Lambda}_{\omega_1}$,
$\mathbf{\Lambda}_{\omega_1} = \mathrm{diag}(\lambda_{\omega_1,1}, ...\lambda_{\omega_1,D})$,
where $\lambda_{\omega_1,1}, ..., \lambda_{\omega_1,D}$ are the eigenvalues of $\mathbf{\Sigma_1}$. From this we can derive eq. 8.

$$\begin{aligned}
\beta_3(\mathbf{v}) &= \mathbf{v}^T\mathbf{B_2}\mathbf{v} - 2\mathbf{b_2^T}\mathbf{v} = t_2 \\
\mathbf{B_2} &= \mathbf{U}_{\omega_1}^\mathbf{T}(\hat{\mathbf{\Sigma}}_\mathbf{1}^{-1} - \hat{\mathbf{\Sigma}}_\mathbf{2}^{-1})\mathbf{U}_{\omega_1} \\
\mathbf{b_2^T} &= (\mu_1 - \mu_2)^T\hat{\mathbf{\Sigma}}_\mathbf{2}^{-1}\mathbf{U}_{\omega_1} \\
\mathbf{v} &\sim N(\mathbf{0}, \mathbf{\Lambda}_{\omega_1})
\end{aligned} \tag{8}$$

2.3. Scale dimensions to normalize all variances in $\mathbf{\Sigma_1}$

We assume that $\mathbf{\Lambda}_{\omega_1}$ is positive-definite and therefore none of the eigenvalues are zero. If some of the eigenvalues are zero, the dimensionality of the problem can either be reduced or the classification problem is trivial (if ω_2 has a variance in this dimension or a different mean). (Of course, an NB classifier may not be responsive to this state of affairs, and therefore perform sub-optimally. However, we do not consider this degenerate special case below.)

We define $\mathbf{u} = \mathbf{\Lambda}_{\omega_1}^{-1/2}\mathbf{v}$ and derive eq. 9.

$$\begin{aligned}
\beta_1(\mathbf{u}) &= \mathbf{u}^T\mathbf{B}\mathbf{u} - 2\mathbf{b_3^T}\mathbf{v} = t_2 \\
\mathbf{B} &= \mathbf{\Lambda}_{\omega_1}^{1/2}\mathbf{U}_{\omega_1}^\mathbf{T}(\hat{\mathbf{\Sigma}}_\mathbf{1}^{-1} - \hat{\mathbf{\Sigma}}_\mathbf{2}^{-1})\mathbf{U}_{\omega_1}\mathbf{\Lambda}_{\omega_1}^{1/2} \\
\mathbf{b_3^T} &= (\mu_1 - \mu_2)^T\hat{\mathbf{\Sigma}}_\mathbf{2}^{-1}\mathbf{U}_{\omega_1}\mathbf{\Lambda}_{\omega_1}^{1/2} \\
\mathbf{u} &\sim N(\mathbf{0}, \mathbf{I})
\end{aligned} \tag{9}$$

2.4. Rotate matrices to diagonalize the quadratic boundary

Now that $\mathbf{u}$ is normally distributed with mean $\mathbf{0}$ and covariance $\mathbf{I}$, it is possible to rotate $\mathbf{B}$ until it is diagonal without inducing any correlation between random variates. Therefore, we define $\mathbf{U_B}$ and $\mathbf{\Lambda_B}$ to be the eigenvector matrix and diagonal eigenvalue matrix of $\mathbf{B}$ respectively.

We finally define $\mathbf{y} = \mathbf{U_B^T}\mathbf{u}$ and derive eq. 10.

$$\begin{aligned}
\beta(\mathbf{y}) &= \mathbf{y}^T\mathbf{\Lambda_B}\mathbf{y} - 2\mathbf{b}^\mathbf{T}\mathbf{y} = t_2 \\
\mathbf{b}^\mathbf{T} &= (\mu_1 - \mu_2)^T\hat{\mathbf{\Sigma}}_\mathbf{2}^{-1}\mathbf{U}_{\omega_1}\mathbf{\Lambda}_{\omega_1}^{1/2}\mathbf{U_B} \\
\mathbf{y} &\sim N(\mathbf{0}, \mathbf{I})
\end{aligned} \tag{10}$$

It is easy to derive the values for $\mathbf{\Phi}$, $\mathbf{m}$ and t in eq. 6 using eq. 10. These values are given in equation 11.

$$\begin{aligned} \alpha_i &= \lambda_{B,i} \quad \forall i \in \{1, ..., D\} \\ m_i &= \frac{b_i}{\lambda_{B,i}} \quad \forall i \in \{1, ..., D\} \\ t &= t_2 + \sum_{i=1}^{D} \frac{b_i^2}{\lambda_{B,i}}. \end{aligned} \tag{11}$$

It is possible for some of the $\lambda_{B,i}$ values to be zero in which case some of the m_i coefficients become infinite or undefined (this is also the case for t). This happens when some of the random variates only have a linear component in eq. 10 or if the variates make no discriminative difference (in which case b_i is also zero). These cases are discussed in the next section.

3. Decision boundaries and their solutions

In this section we discuss all possible quadratic boundaries derivable from the theory developed in section 2. We also give analytical solutions to the error rate performances associated with each decision boundary (except for paraboloidal decision boundaries discussed later).

3.1. Linear decision boundaries

Linear decision boundaries are the simplest case to solve and occur when $\mathbf{\Lambda_B} = \mathbf{B} = \mathbf{0}$. From eq. 9 it is easy to see that $\hat{\Sigma}_1 = \hat{\Sigma}_2$ for this to be true and it follows that

$$\begin{aligned} \int_{\Omega_2} p(\mathbf{x}|\omega_1)d\mathbf{x} &= p(-2\mathbf{b}^{\mathrm{T}}\mathbf{y} > t_2) \\ -2\mathbf{b}^{\mathrm{T}}\mathbf{y} &\sim N(0, 4\mathbf{b}^{\mathrm{T}}\mathbf{b}) \end{aligned} \tag{12}$$

From eq. 12 it is easy to prove that

$$\int_{\Omega_2} p(\mathbf{x}|\omega_1)d\mathbf{x} = \frac{1}{2}\operatorname{erfc}(\frac{t_2}{\sqrt{8\mathbf{b}^{\mathrm{T}}\mathbf{b}}}) \tag{13}$$

3.2. Ellipsoidal decision boundaries

Ellipsoidal decision boundaries occur when either $\mathbf{B}$ or $-\mathbf{B}$ is positive-definite. In other words the eigenvalues $\lambda_{B,1}, ..., \lambda_{B,D}$ are either all negative or all positive. This is a special case that occurs in NB classifiers when one class consistently has a larger variance than the other class for all dimensions. Since $\mathbf{m}$ (see eq. 11) is defined (none of the eigenvalues are zero), we can attempt to solve eq. 6. Many solutions have been proposed for this problem (see, for example [14]), but the one that we find most efficient is proposed in [13, 15] and is restated here.

Theorem 1. For $\mathbf{y} \sim N(\mathbf{0}, \mathbf{I})$ and $F(\Phi, \mathbf{m}, t)$ as defined in eq. 6, we have

$$F(\Phi, \mathbf{m}, t) = \sum_{i=0}^{\infty} \alpha_i F_{D+2i}(\frac{t}{p}), \quad \phi_i > 0 \ \ \forall i \in \{1, ..., D\},$$

where $F_n(x)$ is defined to be the cdf of a central chi-square distribution with n degrees of freedom, p is any constant satisfying

$$0 < p \leq \phi_i \quad \forall i \in \{1, ..., D\},$$

and α_i can be calculated with the recurrence relations

$$\begin{aligned} \alpha_0 &= \exp(-\frac{1}{2}\sum_{j=1}^{D} m_j^2)\sqrt{\prod_{j=1}^{D} p/\phi_j} \\ \alpha_i &= \frac{1}{2i}\sum_{j=0}^{i-1} \alpha_j g_{i-j} \\ g_r &= \sum_{i=1}^{D}(1 - p/\phi_i)^r + rp\sum_{i=1}^{D} \frac{m_i^2}{\phi_i}(1 - p/\phi_i)^{r-1} \end{aligned}$$

Also, the α coefficients above will always converge and

$$\sum_{i=0}^{\infty} \alpha_i = 1$$

Finally, a bound can be placed on the error from summing only k terms as follows

$$\begin{aligned} 0 &\leq F(\Phi, \mathbf{m}, t) - \sum_{i=0}^{k-1} \alpha_i F_{D+2i}(\frac{t}{p}) \\ &\leq (1 - \sum_{i=1}^{k-1} \alpha_i) F_{D+2k}(t/p) \end{aligned}$$

Proof. The proof can be found in [15].

For optimal convergence in the above series we select $p = \inf\{\phi_1, ..., \phi_D\}$, the largest possible value for p.

A useful recurrence relation for calculating $F_n(x)$ is as follows

$$\begin{aligned} F_1(x) &= \operatorname{erf}(\sqrt{\frac{x}{2}}) \\ F_2(x) &= 1 - \exp(-\frac{x}{2}) \\ F_{n+2}(x) &= F_n(x) - \frac{(x/2)^{n/2}e^{-x/2}}{\Gamma(n/2+1)} \end{aligned} \tag{14}$$

We discussed analytical solutions for the case where all α_i's are greater than zero. A symmetric statement can be made for all α_i's less than zero. Therefore, we conclude that

$$\begin{aligned} &\int_{\Omega_2} p(\mathbf{x}|\omega_1)d\mathbf{x} \\ &= \begin{cases} F(-\Phi, \mathbf{m}, -t) & \sup\{\phi_1, ..., \phi_D\} < 0 \\ 1 - F(\Phi, \mathbf{m}, t) & \inf\{\phi_1, ..., \phi_D\} > 0 \end{cases} \end{aligned} \tag{15}$$

3.3. Hyperboloidal decision boundaries

Hyperboloidal decision boundaries occur when $\mathbf{B}$ is indefinite and invertible. Therefore, some of the eigenvalues of $\mathbf{B}$ will be positive and others negative, but none of them zero. This is the most frequently occurring case and also the most difficult to solve. Although much research has been done on solving the definite quadratic form (as for the elliptic boundary discussed above), finding an exact analytical expression for the indefinite quadratic form has been unsuccessful (see [12, 13, 14, 16]). The existing solutions all lead to estimates, bounds or unwieldy solutions (and unusable for NB error analysis). In contrast, we propose a solution that is exact and efficient.

Theorem 2. For $\mathbf{y} \sim N(\mathbf{0}, \mathbf{I})$ and $F(\Phi, \mathbf{m}, t)$ as defined in eq. 9, we can rewrite $F(\Phi, \mathbf{m}, t)$ as follows.

$$\begin{aligned} F(\Phi, \mathbf{m}, t) &= p(\sum_{i=1}^{d_1} \phi_i'(y_i - m_i')^2 - \\ &\quad \sum_{j=1}^{d_2} \phi_j^*(y_{d_1+j} - m_j^*)^2 \leq t), \\ \phi_i', \phi_j^* &> 0 \qquad \forall i \in \{1, ..., d_1\}, \forall j \in \{1, ..., d_2\}, \end{aligned}$$

where $d_1 + d_2 = D$. From this, we can show that

$$F(\Phi, \mathbf{m}, t) = 1 - \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \alpha_i' \alpha_j^* \Upsilon_{d_1+2i, d_2+2j}(t/p), \quad t \geq 0$$

where we calculate the α_i' and α_j^* coefficients by applying theorem 1 (with common value p) to $F(\Phi', \mathbf{m}', t)$ and $F(\Phi^*, \mathbf{m}^*, t)$ respectively. Note that the α_i' and α_j^* coefficients are independent of t. p can be any arbitrary constant satisfying

$$0 < p \leq \phi_i', \phi_j^* \quad \forall i \in \{1, ..., d_1\}, \forall j \in \{1, ..., d_2\}$$

$\Upsilon_{k_1,k_2}(z)$ can be calculated using the following recurrence relations.

$$\begin{aligned} \Upsilon_{-1,0}(z) &= \frac{1}{\sqrt{\pi}} \Gamma(1/2, z/2) \\ \Upsilon_{-1,1}(z) &= \frac{1}{2}\left[1 - \frac{z}{2}\left(K_0(\frac{z}{2})\mathbf{L}_{-1}(\frac{z}{2}) + K_1(\frac{z}{2})\mathbf{L}_0(\frac{z}{2})\right)\right] \\ \Upsilon_{0,k_2}(z) &= 2^{-k_2/2} e^{-z/2} \\ \Upsilon_{k_1,k_2}(z) &= \Upsilon_{k_1-2,k_2}(z) + D_{k_1,k_2}(z) \\ \Upsilon_{k_1,k_2}(z) &= \Upsilon_{k_1,k_2-2}(z) - D_{k_1,k_2}(z), \end{aligned}$$

where

$$D_{k_1,k_2}(z) = \frac{e^{-z/2}}{2^{(k_1+k_2)/2-1}\Gamma(k_1/2)} \psi(1 - \frac{k_1}{2}, 2 - \frac{k_1+k_2}{2}; z)$$

$\Gamma(a)$ is the gamma function and $\Gamma(a, x)$ is the upper incomplete gamma function. $K_n(x)$ is the modified Bessel function of the second kind and $\mathbf{L}_n(x)$ is the modified Struve function. $\psi(a, b; z)$ is the Tricomi confluent hypergeometric function (also known as the $U(a, b; z)$ function discussed in [17]).

Finally, a bound can be placed on the error from summing only K and L terms.

$$\begin{aligned} 0 &< 1 - \sum_{i=0}^{K} \sum_{j=0}^{L} \alpha_i' \alpha_j^* \Upsilon_{d_1+2i, d_2+2j}(t/p) - F(\Phi, \mathbf{m}, t) \\ &\leq (1 - \sum_{i=0}^{K-1} \alpha_i')(\sum_{j=0}^{L-1} \alpha_j^*) \Upsilon_{d_1+2K, d_2+2L}(t/p) \\ &\quad + 1 - \sum_{j=0}^{L-1} \alpha_j^* \end{aligned}$$

Proof. Partial proofs can be found in [12, 13]. Unfortunately, the full proof of this theorem is fairly involved and will be provided in a future paper.

It becomes impractical to calculate $D_{k_1,k_2}(z)$ for large values of k_1 and k_2 and therefore the following recurrence relations become useful

$$\begin{aligned} D_{k_1,k_2}(z) &= \frac{1}{4 - 2k_1}[(4 - k_1 - k_2 - 2z)D_{k_1-2,k_2}(z) \\ &\quad + zD_{k_1-4,k_2}(z)] \\ D_{k_1,k_2}(z) &= \frac{1}{4 - 2k_2}[(4 - k_1 - k_2 + 2z)D_{k_1,k_2-2}(z) \\ &\quad - zD_{k_1,k_2-4}(z)] \\ D_{k_1,k_2}(z) &= \frac{1}{2}(D_{k_1-2,k_2}(z) + D_{k_1,k_2-2}(z)) \end{aligned} \tag{16}$$

Although it is theoretically possible to use only the first two recurrence relations in eq. 16, numerical experiments show that when combined, quantization noise will increase rapidly with each iteration. Therefore we use the first two recurrence relations independently and fill all the remaining gaps with recurrence relation three in eq. 16. Notice that theorem 2 only applies for cases where $t \geq 0$. A symmetric argument can be expressed for cases where $t < 0$. Finally, we conclude that

$$\begin{aligned} &\int_{\Omega_2} p(\mathbf{x}|\omega_1) d\mathbf{x} \\ &= \begin{cases} F(-\Phi, \mathbf{m}, -t) & t < 0 \\ 1 - F(\Phi, \mathbf{m}, t) & t \geq 0 \end{cases} \end{aligned} \tag{17}$$

3.4. Cylindrical decision boundaries

Cylindrical decision boundaries occur when some of the eigenvalues $\lambda_{B,i}$ and their corresponding linear parts b_i are zero. It is fairly easy to see from eq. 10 that these features can simply be dropped and the dimensionality decreased.

3.5. Paraboloidal decision boundaries

Paraboloidal decision boundaries occur when some of the eigenvalues $\lambda_{B,i}$ are zero, but their corresponding linear parts b_i are none-zero. In the context of NB classifiers, this only happens when some of the estimated variances (in a given dimension) are identical for ω_1 and ω_2, but their means differ. Unfortunately, an exact solution for this problem does not yet exist. Therefore, as a temporary solution, we simply add a small disturbance $\delta\lambda_i$ to eq. 10 to get an approximate hyperboloidal or ellipsoidal decision boundary.

4. Results

In this section, we compare the error performance of simple binary classifiers of different dimensionalities for both the Bayes error rate and that obtained using NB classifiers. These error rates will be obtained using two methods: Monte-Carlo simulations and the analytical methods proposed above. Our experimental configurations are similar to those proposed in [13].

4.1. Example 1: A two dimensional classification problem

For this example we will explore the error rates of a two dimensional Gaussian binary classification problem with parameters

$$\mu_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \Sigma_1 = \alpha \begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix},$$

$$\mu_2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix} \quad \Sigma_2 = \alpha \begin{bmatrix} 5 & -2 \\ -2 & 1 \end{bmatrix},$$

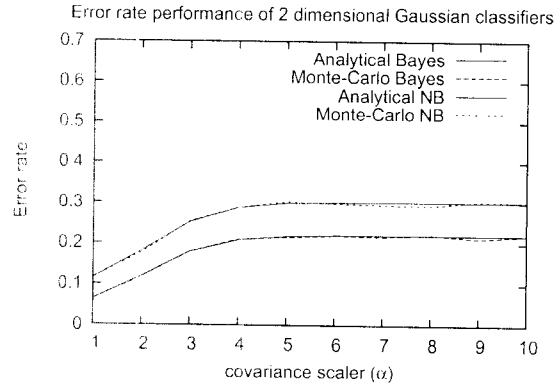

Figure 1: *Naive and Bayes error rates for two dimensional problem in example 1 with increasing class covariances.*

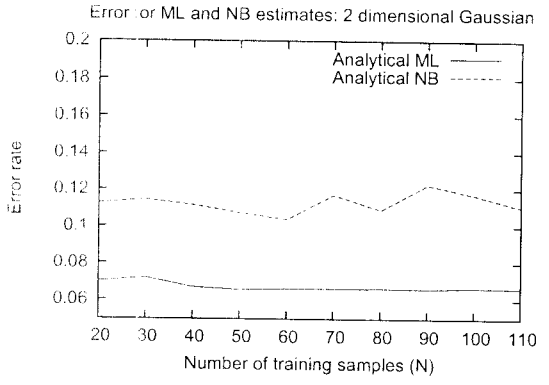

Figure 2: *Naive and maximum likelihood estimate error rates for two dimensional problem in example 2 while increasing the number of training samples.*

where α is a covariance scalar. Figure 1 shows the Bayes and NB (perfect estimate) error rates obtained with the analytical model developed and Monte-Carlo simulations. For this experiment $p(\omega_1) = p(\omega_2) = 0.5$ and 10000 samples in total were generated for the simulations.

Figure 2 shows the analytical results obtained for $\alpha = 1$ where we estimate both the Maximum likelihood (ML) and NB parameters using a varying number of training samples.

It is clear from this experiment that the low dimensional ML classifier provides superior performance to the NB classifier, and that our analytic estimates agree with those obtained by Monte-Carlo simulation.

4.2. Example 2 : A 12 dimensional classification problem

Now we explore a high dimensional problem (12 dimensional) to illustrate the power of NB classifiers. For this example we

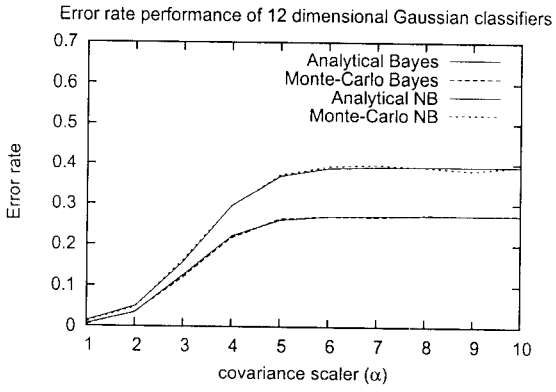

Figure 3: *Naive and Bayes error rates for 12 dimensional problem in example 2 with increasing class covariances.*

define

$$\mu_1 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 1 \end{bmatrix} \quad \Sigma_1 = \alpha \begin{bmatrix} 5 & -1 & 0 & \dots & 0 \\ -1 & 5 & -1 & & 0 \\ \vdots & \ddots & \ddots & \ddots & \\ 0 & & -1 & 5 & -1 \\ 0 & \dots & 0 & -1 & 5 \end{bmatrix},$$

$$\mu_2 = \begin{bmatrix} -1 \\ -1 \\ \vdots \\ -1 \\ -1 \end{bmatrix} \quad \Sigma_2 = \alpha \begin{bmatrix} 6 & -2 & 0 & \dots & 0 \\ -2 & 4 & -2 & & 0 \\ \vdots & \ddots & \ddots & \ddots & \\ 0 & & -2 & 6 & -2 \\ 0 & \dots & 0 & -2 & 4 \end{bmatrix},$$

where α is a covariance scalar. Figure 3 shows the Bayes and NB (perfect estimate) error rates obtained with the analytical model developed and Monte-Carlo simulations. For this experiment $p(\omega_1) = p(\omega_2) = 0.5$ and 10000 samples in total were generated for the simulations.

Figure 4 shows the analytical results obtained for $\alpha = 1$ where we estimate both the Maximum likelihood (ML) and NB parameters using a varying number of training samples.

It is clear from figure 4 that for high dimensional problems, NB classifiers perform better when data sparsity is an issue. This is due to the high variance in the ML estimate. NB classifiers are robust for sparse problems and for this specific problem, NB performs relatively well even when more than a hundred training samples are provided.

5. Conclusion

In this paper, we derived analytical solutions for calculating error probabilities in correlated Gaussian feature spaces for arbitrary quadratic decision boundaries. We applied the theory in the context of NB classifiers and showed the validity for both a 2 and 12 dimensional problem by comparing the analytical solutions to those obtained with Monte-Carlo simulations. Both of these case-studies had hyperboloidal Bayes and NB decision

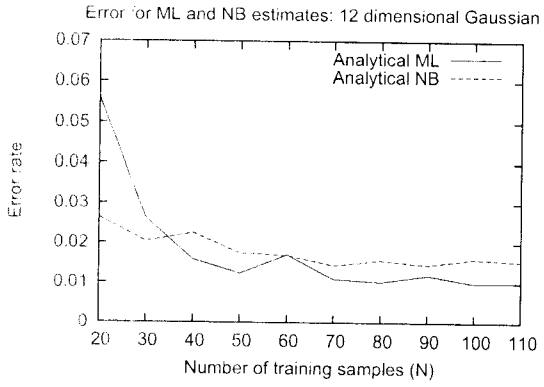

Figure 4: *Naive and maximum likelihood estimate error rates for 12 dimensional problem in example 2 while increasing the number of training samples.*

boundaries, a problem that had not been solved analytically previously.

We also demonstrated the robust behavior of NB classifiers in data sparse and high dimensional environments (see figure 4).

Unfortunately, we still don't have a proper solution for the paraboloidal decision boundaries and we suggested a method for approximating the boundary with a hyperboloidal or ellipsoidal boundary; this method has also been proposed in [13]. It should be noted that this method is not without problems, since the α_i terms in theorem 1 take longer to converge when an exceptionally small ϕ_i value or large m_i value is present. From eq. (11) it is clear that a small value for $\lambda_{B,i}$ will produce a small value for α_i and a large value for m_i.

For future work, we propose to find an exact analytical solution for the error rates obtained when paraboloidal decision boundaries occur. Although these boundaries are themselves degenerate (requiring exactly equal class covariances), the same computational issues arise when the hyperboloidal boundaries are almost paraboloidal (i.e. when the relevant class covariances are close).

6. References

[1] S. J. Russel and P. Norvig, *Artificial Intelligence: a Modern Approach*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1995.

[2] E. van Dyk and E. Barnard, "Naive bayesian classifiers for multinomial features: a theoretical analysis," in *Proceedings of the 18th Annual Symposium of the Pattern Recognition Association of South Africa*, South Africa, 2007, pp. 75–82.

[3] D.J. Hand and K.Yu, "Idiot bayes ? not so stupid after all?," *International Statistical Review*, vol. 69, no. 3, pp. 385–399, 2001.

[4] P. Domingos and M.Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine-Learning*, vol. 29, pp. 103–130, 1997.

[5] A. Webb, *Statistical Pattern Recognition*, John Wiley & Sons, Ltd., England, second edition, 2002.

[6] H. Chernoff, "A measure for asymptotic efficiency of a hypothesis based on a sum of observations," *The Annels of Mathematical Statistics*, vol. 23, pp. 493–507, 1952.

[7] T. Ito, "Approximate error bounds in pattern recognition," *Machine Intelligence*, vol. 7, pp. 369–372, 1972.

[8] M. Hellman, "Probability of error, equivocation, and chernoff bound," *IEEE Transactions on Information Theory*, vol. 16, pp. 368–372, 1970.

[9] P. Deijver, "On a new class of bounds on bayes risk in multihypthesis pattern recognition," *IEEE Transactions on Computers*, vol. 23, pp. 70–80, 1974.

[10] W. Hashlamoun, P. Varshney, and V. Samarasooriya, "A tight upper bound on the bayesian probability of error," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 2, pp. 220–225, 1994.

[11] H. Avi-Itzhak and T. Diep, "Arbitrarily tight upper and lower bounds on the bayesian probability of error," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 1, pp. 89–91, 1996.

[12] S.J. Press, "Linear combinations of non-central chi-square variates," *The Annels of Mathematical Statistics*, vol. 37, no. 2, pp. 480–487, 1966.

[13] M.H El Ayadi, M.S. Kamel, and F. Karray, "Toward a tight upper bound for the error probability of the binary gaussian classification problem," *Pattern Recognition*, vol. 41, pp. 2120–2132, 2008.

[14] B. Shah, "Distribution of definite and of indefinite quadratic forms from a non-central normal distribution," *The Annels of Mathematical Statistics*, vol. 34, no. 1, pp. 186–190, 1963.

[15] H. Ruben, "Probability content of regions under spherical normal distributions, iv: the distribution of homogeneous and non-homogeneous quadratic functions of normal variables," *The Annels of Mathematical Statistics*, vol. 33, no. 2, pp. 542–570, 1962.

[16] D. Raphaeli, "Distribution of noncentral quadratic forms in complex normal variables," *IEEE Transactions on Information Theory*, vol. 42, no. 3, pp. 1002–1007, 1996.

[17] L.J. Slatêr, *Confluent Hypergeometric Functions*, Cambridge University. Press, London, 1960.