CATÓLICA
UNIVERSIDADE CATÓLICA PORTUGUESA | PORTO
Faculdade de Economia e Gestão

DOCUMENTOS DE TRABALHO WORKING PAPERS

GESTÃO MANAGEMENT

Nº 09/2009

LINEAR DISCRIMINANT RULES FOR HIGH-DIMENSIONAL CORRELATED DATA: ASYMPTOTIC AND FINITE SAMPLE RESULTS

Pedro Duarte Silva

Universidade Católica Portuguesa (Porto)

Linear discriminant rules for high-dimensional correlated data: Asymptotic and finite sample results

A. Pedro Duarte Silva

Faculdade de Economia e Gestão & CEGE, Univ. Católica Portuguesa at Porto,
Rua Diogo Botelho, 1327, 4169-005 Porto, Portugal; psilva@porto.ucp.pt

Abstract. A new class of linear discrimination rules, designed for problems with many correlated variables, is proposed. This proposal tries to incorporate the most important patterns revealed by the empirical correlations and accurately approximate the optimal Bayes rule as the number of variables increases. In order to achieve this goal, the new rules rely on covariance matrix estimates derived from Gaussian factor models with small intrinsic dimensionality.

Asymptotic results, based on a analysis that allows the number of variables to grow faster than the number of observations, show that the worst possible expected error rate of the proposed rules converges to the error of the optimal Bayes rule when the postulated model is true, and to a slightly larger constant when this model is a close approximation to the data generating process.

Simulation results suggest that, in the data conditions they were designed for, the new rules can clearly outperform both Fisher's and naive linear discriminant rules.

Key words: Discriminant Analysis, High Dimensionality, Expected Misclassification Rate, Min-Max Regret.

AMS2000 subject classification: 62H30

1 Introduction

The classical theory of Linear Discriminant Analysis (see, for example, [7]) assumes the existence of a training data set with more observations than variables leading to a non-singular empirical covariance matrix. However, nowadays many applications work with data bases where a large number of variables is measured on a smaller set of observations. Practical experience has shown [2, 3] that, for problems of this type, natural extensions of Fisher's linear discriminant rule, that replace the inverse of the empirical covariance matrix by a generalized inverse, have a disappointing performance. On the other hand, in the same problems the naive discriminant rule that ignores all variable correlations can be quite effective.

Recently, Bickel and Levina [1], based on an asymptotic analysis that allows the number of variables to grow faster than the number of observations, have shown that these surprising results have a deep theoretical justification, and that the expected error of the naive rule can approach a constant close to the expected error of the optimal Bayes rule, while generalized versions of Fisher's rule are asymptotically no better than simple random guessing, ignoring the data.

Here, it will be shown that linear discriminant rules based on covariance estimates derived from low-dimensional factor models can successfully incorporate some of the information available on the empirical correlations, and under conditions similar to those considered in [1] can achieve, or come close to, asymptotic optimality for some problems where both Fisher's and naive Bayes rules perform poorly.

The reminder of this paper is organized as follows. Section 2 motivates and introduces our proposal. Section 3 presents its asymptotic properties. Section 4 addresses implementation issues. Section 5 describes preliminary simulation experiments assessing the performance of the new rules in finite samples. Section 6 discusses the main results of this contribution and presents perspectives for future research. Mathematical proofs are given in the Appendix Section A.

2 A Factor Model Linear Discriminant Rule

Consider the two-group homocedastic Gaussian model where entities are represented by binary pairs $(X, Y); X \in \Re^p; Y \in \{0, 1\}$ and the distribution of X conditioned on Y is the multivariate normal $N_p(\mu_{(Y)}, \Sigma)$. The classical discriminant problem deals with the development of rules capable of predicting unknown Y values (class lables) given X observations. When the parameters $\mu_{(0)}, \mu_{(1)}, \Sigma$ are known and a-priori probabilities $\pi_0 = P(Y = 0), \pi_1 = P(Y = 1)$ are equal (i.e., $\pi_0 = \pi_1 = 1/2$) it is well known [7] that the classification rule that minimizes the expected misclassification error is the theoretical Bayes rule, given by

$$Y = \delta_B(X) = \mathbf{1}(\Delta^T \Sigma^{-1} \gamma > 0) \qquad (1)$$

where $\Delta = \mu_{(1)} - \mu_{(0)}$; $\gamma = X - \frac{1}{2}(\mu_{(0)} + \mu_{(1)})$ and $\mathbf{1}(.)$ is the indicator function.

Different a-priori probabilities and/or misclassification costs can be easily incorporated into (1), and would not alter the essential of the arguments made here, but for the sake of simplicity will be omitted from this paper.

In practice, Δ, γ and Σ are usually unknown and are estimated from a training sample of $n = n_0 + n_1$ observations $((X_i, Y_i)$; $i = 1, ..., n)$ with known class labels. When $n_0 >> p$ and $n_1 >> p$ common estimators are

$$\hat{\Delta}_F = \bar{X}_1 - \bar{X}_0 = \frac{1}{n_1} \sum_{Y_i=1} X_i - \frac{1}{n_0} \sum_{Y_i=0} X_i$$

$$\hat{\gamma}_F = X - \frac{1}{2} \left[\bar{X}_0 + \bar{X}_1 \right]$$

$$\hat{\Sigma}_F = \frac{1}{n-2} \left[\sum_{Y_i=0} (X_i - \bar{X}_0)(X_i - \bar{X}_0)^T + \sum_{Y_i=1} (X_i - \bar{X}_1)(X_i - \bar{X}_1)^T \right]$$

which, for non-singular $\hat{\Sigma}_F$, leads to Fisher's linear rule

$$Y = \delta_F(X) = \mathbf{1}(\hat{\Delta}_F^T \ \hat{\Sigma}_F^{-1} \hat{\gamma}_F > 0) \tag{2}$$

Here, we will be mostly concerned with problems where p is close to, or higher than n. In the later case $\hat{\Sigma}_F$ is singular and rule (2) can not be applied directly. However, a modified Fisher's rule

$$Y = \delta_{MF}(X) = \mathbf{1}(\hat{\Delta}_F^T \ \hat{\Sigma}_F^{-} \hat{\gamma}_F > 0) \tag{3}$$

can be defined by replacing $\hat{\Sigma}_F^{-1}$ by $\hat{\Sigma}_F$ Moore-Penrose generalized inverse

$$\hat{\Sigma}_F^{-} = \frac{1}{k} \sum_{a=1}^{k} \frac{1}{\hat{\lambda}_a} \hat{\xi}_a \hat{\xi}_a^{\,T}$$

where k is the rank of $\hat{\Sigma}_F$ and $\hat{\lambda}_a$, $\hat{\xi}_a$ their non-null eigenvalues and corresponding normalized eigenvectors.

Alternatively, when Σ is estimated by the diagonal matrix of training sample variances, $\hat{\Sigma}_I = \text{diag}(\hat{\Sigma}_F)$, one gets an estimator of the optimal rule for independent variables, known as naive Bayes

$$Y = \delta_I(X) = \mathbf{1}(\hat{\Delta}_F^T \ \hat{\Sigma}_I^{-1} \hat{\gamma}_F > 0) \tag{4}$$

Note that, contrary to $\hat{\Sigma}_F$, $\hat{\Sigma}_I$ is always non-singular for any value of n and p, as long as the empirical variances of X remain strictly positive. Furthermore, several authors (e.g. [2, 3]) remark that rule (4) is surprisingly effective even in problems where many variables are clearly correlated. However, when $p >> n$ some form of regularization or variable selection may be necessary to avoid error accumulation in the estimation of Δ and γ (see [6] and [4]).

Recently, Bickel and Levina [1] have shown that under general conditions where both p and n grow to infinity and $n/p \to d < \infty$, a variant of rule (4) that replaces $\hat{\Delta}_F$ and $\hat{\gamma}_F$ by consistent (when $p \to \infty$) estimators of Δ and γ, has an asymptotic error rate that depends on the maximum ratio between the largest and the smallest eigenvalues of population correlation matrices. When this ratio is bounded by some "moderate" constant the asymptotic performance of the naive rule is close to that of the optimal Bayes rule.

Building on these results, here we will propose an alternative linear rule with good asymptotic performance for some common data conditions where ratios of correlation eigenvalues can be large. In particular, we will assume that the true population covariance can be reasonably approximated by a covariance matrix derived from the following q-dimensional ($q << p$) factor model

$$
\begin{aligned}
X = \mu_{(Y)} + \beta\ F + \Omega\ \epsilon \\
\beta \in \Re^{p*q}\ ;\ a > j\ \Rightarrow \beta(j,a) = 0 \\
\Omega \in \Re^{p*p}\ ;\ j \neq l\ \Rightarrow \Omega(j,l) = 0\ ;\ \Omega(j,j) > k_0 \in \Re_+
\end{aligned}
\quad (5)
$$

where F and ϵ are respectively q-dimensional and p-dimensional random vectors following $N_q(0, I_q)$ and $N_p(0, I_p)$ distributions, and the condition $a > j\ \Rightarrow \beta(j,a) = 0$, is imposed to ensure the identifiability of β.

When model (5) holds the X covariance matrix, given by $\Sigma = \beta\beta^T + \Omega^2$, is non-singular with inverse equal to

$$\Sigma^{-1} = \Omega^{-2} - \Omega^{-2}\beta[I_q + \beta^T\Omega^{-2}\beta]^{-1}\beta^T\Omega^{-2} \quad (6)$$

This suggests the rule

$$Y = \delta_{Fct_q}(X) = \mathbf{1}(\hat{\Delta}_C^T\ \hat{\Sigma}_{Fct_q}^{-1}\hat{\gamma}_C > 0) \quad (7)$$

where $\hat{\Delta}_C$ and $\hat{\gamma}_C$ are consistent estimators of Δ and γ, to be discussed later, and $\hat{\Sigma}_{Fct_q}$ is given by

$$\hat{\Sigma}_{Fct_q} = \hat{\beta}\ \hat{\beta}^T + \hat{\Omega}^2\ ;\ (\hat{\beta}, \hat{\Omega}) = argmin||\hat{\Sigma}_{Fct_q} - \hat{\Sigma_F}||^2 \quad (8)$$

with $||.||$ being the Frobenius matrix norm, $||A||^2 = tr(A^T A) = \sum_{j,l} A(j,l)^2$.

3 Asymptotic Properties

In this section we will discuss the min-max asymptotic performance of rule δ_{Fct_q} when $n, p \to \infty$ and $n/p \to d < \infty$. In particular, we will be concerned with the conditions for convergence, and the limit, of the min-max expected misclassification error

$$\overline{W}_{\Gamma_{Fct_q}}(\delta_{Fct_q}) = max_{\Gamma_{Fct_q}}\left[P_\theta(\delta_{Fct_q}(X) = 1|Y = 0)\right] \quad (9)$$

where

$$\theta = (\Delta, \gamma, \Sigma) \in \Gamma_{Fct_q}(k_0, k_1, k_2, q, B, c)$$

and

$$\mathit{\Gamma}_{Fct_q}(k_0,k_1,k_2,q,B,c)=\left\{\begin{array}{c}(\mathit{\Delta},\gamma,\mathit{\Sigma}): \\ \mathit{\Delta}^T\mathit{\Sigma}^{-1}\mathit{\Delta}\geq c^2 \\ k_1\leq\lambda_{min}(\mathit{\Sigma})\leq\lambda_{max}(\mathit{\Sigma})\leq k_2 \\ \mathit{\Delta},\gamma\in B \\ \forall j=1,...,p\ ;\ a=1,...,q \\ \sum_{j',l'}\left|\frac{\partial\beta(j,a)}{\partial\mathit{\Sigma}(j',l')}\right|\rightarrow e<\infty \\ \sum_{j',l'}\left|\frac{\partial\mathit{\Omega}(j,j)}{\partial\mathit{\Sigma}(j',l')}\right|\rightarrow f<\infty\end{array}\right\}$$

$$\begin{array}{l}\mathit{\Sigma}_{Fct_q}=\beta\ \beta^T+\mathit{\Omega}^2\ ;\ (\beta,\mathit{\Omega})=argmin||\ \mathit{\Sigma}_{Fct_q}-\mathit{\Sigma}||^2 \\ \qquad\beta\in\Re^{p*q}\ ;\ a>j\ \Rightarrow\beta(j,a)=0 \\ \qquad\mathit{\Omega}\in\Re^{p*p}\ ;\ j\neq l\ \Rightarrow\mathit{\Omega}(j,l)=0\ ;\ \mathit{\Omega}(j,j)>k_0\in\Re_+\end{array}$$

with $\lambda_{min}(\mathit{\Sigma})$ and $\lambda_{max}(\mathit{\Sigma})$ being the smallest and largest eigenvalues of $\mathit{\Sigma}$, and B a compact subset of $l_2(N)$, the set of real number sequences with convergent square sums.

The specification of the parameter space, $\mathit{\Gamma}_{Fct_q}$, requires some explanation.

The condition $\mathit{\Delta}^T\mathit{\Sigma}^{-1}\mathit{\Delta}\geq c^2$ establishes the minimum degree of group separation on $\mathit{\Gamma}_{Fct_q}$. It is well known that for fixed θ, the expected error rate of the optimal Bayes rule equals $1-\mathit{\Phi}(\frac{1}{2}\sqrt{\mathit{\Delta}^T\mathit{\Sigma}^{-1}\mathit{\Delta}})$ with $\mathit{\Phi}(.)$ being the cumulative probability of a standardized Gaussian random variable. Therefore, this condition implies that for all $\theta\in\mathit{\Gamma}_{Fct_q}$ the optimal misclassification rate is bounded from above by $1-\mathit{\Phi}(c/2)$, which then becomes an useful benchmark against which the asymptotic rate of any empirical rule can be assessed.

Condition $k_1\leq\lambda_{min}(\mathit{\Sigma})\leq\lambda_{max}(\mathit{\Sigma})\leq k_2$ ensures that $\mathit{\Sigma}$ is always nonsingular and well-conditioned.

Conditions $\mathit{\Delta},\gamma\in B$ are technical requirements necessary to allow the possibility of estimating $\mathit{\Delta}$ and γ consistently. We note that when $\mathit{\Delta}\notin l_2(N)$ and $||\mathit{\Sigma}||$ is bounded, the expected misclassification rate of the Bayes rule converges to zero when p grows without limit. In that case it may be possible to find empirical rules with similar perfect asymptotic performance, even if their coefficients remain far apart from those of the theoretical rule. While such problems may have some interest on their own, they will be not considered here, and we will focus on the more standard conditions where rules approaching perfect group separation are not possible. Therefore, we assume that $\mathit{\Delta}\in l_2(N)$ and look for $\mathit{\Delta},\gamma$ estimators such that $E_\theta||\hat{\mathit{\Delta}}-\mathit{\Delta}||^2=o(1)$ and $E_\theta||\hat{\gamma}-\gamma||^2=o(1)$. Known results in the theory of countable Gaussian sequences (see [6] and Lemma 1 in [1]) show that such estimators exist if and only if $\mathit{\Delta}$ and γ are restricted to lie on a compact subset of $l_2(N)$.

The previous conditions are similar to corresponding conditions assumed by Bikel and Levina [1] in their theoretical study of the naive rule.

Conditions

$$\forall j,a \quad \sum_{j',l'} \left| \frac{\partial \beta(j,a)}{\partial \Sigma(j',l')} \right| \rightarrow e < \infty \quad \sum_{j',l'} \left| \frac{\partial \Omega(j,j)}{\partial \Sigma(j',l')} \right| \rightarrow f < \infty \tag{10}$$

are new technical requirements, specific to the δ_{Fct_q} rule, that are necessary to ensure that convergence of $\hat{\Sigma}_F$ to Σ leads to convergence of $\hat{\Sigma}_{Fct_q}$ to Σ_{Fct_q}. In practice, they imply that any variable (X_j) contribution to the underlying structure of the closest q-factor model is stable, and can be essentially recovered after a finite number of new variables are added to the model. This seems to be a sensible and reasonable assumption, were it not true and no stable q-factor model could be used to approximate the covariance structure defined by the sequence of X variables.

Condition

$$\forall j \quad \Omega(j,j) > k_0 \in \Re_+ \tag{11}$$

ensures that for $(\Delta, \gamma, \Sigma) \in \Gamma_{Fct_q}, \Sigma_{Fct_q}$ remains always non-singular and well-conditioned. The empirical versions of this condition and formula (6) are central in guaranteeing that, similarly to $\hat{\Sigma}_I$ and unlike $\hat{\Sigma}_F$, $\hat{\Sigma}_{Fct_q}$ can always be inverted and leads to an inverse approximation error, $||\hat{\Sigma}_{Fct_q}^{-1} - \Sigma_{Fct_q}^{-1}||$, that can be bounded by a constant times the $||\hat{\Sigma}_{Fct_q} - \Sigma_{Fct_q}||$ error. Furthermore, formula (6) shows that in order to compute $\hat{\Sigma}_{Fct_q}^{-1}$ only a q-dimensional matrix needs to be explicitly inverted, a fact that makes the implementation δ_{Fct_q} computationally feasible for moderately large values of p as long as, as implied by our model, q remains much smaller than p.

We can now state the main result of this section.

Theorem 1.

When $(ln\ p^2)/n \rightarrow 0$,

$$lim\ sup_{n\rightarrow\infty}\ \overline{W}_{\Gamma_{Fct_q}}(\delta_{Fct_q}) \leq 1 - \Phi\left(\frac{\sqrt{K_{0F_q}}}{1+K_{0F_q}}\ c \right)$$

where

$$K_{0F_q} = max_{\Gamma_{Fct_q}} \frac{\lambda_{max}(\Sigma_{0Fct_q})}{\lambda_{min}(\Sigma_{0Fct_q})}\ ;\ \Sigma_{0Fct_q} = \Sigma_{Fct_q}^{-\frac{1}{2}}\ \Sigma\ \left(\Sigma_{Fct_q}^{-\frac{1}{2}}\right)^T$$

and $\Sigma_{Fct_q}^{-\frac{1}{2}}$ is the inverse of the lower-triangular Cholesky decomposition of Σ_{Fct_q}.

For the proof see section A.

Note that the bound defined in Theorem 1 has the same form as the limit found in [1] for the min-max expected error of the naive rule, but replaces the bound (K_0) on the ratio for the eigenvalues of the correlation matrix by K_{0F_q}. This constant measures the maximum distance between the true covariance and a covariance compatible with the postulated q-factor model. When the data generating process satisfies (5), Σ_{0Fct_q} becomes the p-dimensional identity, K_{0F_q} equals one, and the worst expected error rate of δ_{Fct_q} converges to the expected error rate of the optimal rule. On the other hand, when K_{0F_q} is allowed to increase without limit as p grows, the true generating process becomes farther and farther apart from the postulated model and rule δ_{Fct_q} is asymptotically no better than simple random guessing. In intermediate cases, with $K_{0F_q} > 1$ but bounded by some finite constant, rule δ_{Fct_q} does not converge to the theoretical Bayes rule but, depending on the particular value of K_{0F_q}, can be close.

The main motivation for our proposal is the fact that, when the data generating process implies a correlation structure that is far from total independence but close to a low dimensional factor model, K_{0F_q} can be much smaller than K_0. In such a case, Theorem 1 shows that as p grows δ_{Fct_q} can approach a smaller expected error rate than δ_I. The simulation results presented in section 5 suggest that for these conditions, δ_{Fct_q} can perform considerably better than δ_I, δ_F or δ_{MF}, also for moderate values of p and n.

4 Shrinkage and Variable Selection

In order to implement rule δ_{Fct_q}, one has to chose appropriate estimators for Δ and γ. The asymptotic properties described in the previous section are valid for any estimators such that $E_\theta||\hat{\Delta} - \Delta||^2 = o(1)$ and $E_\theta||\hat{\gamma} - \gamma||^2 = o(1)$, and these conditions are satisfied by all shrunken, or truncated, linear estimators of the form

$$\hat{\Delta}_C = c^T \hat{\Delta}_F \ ; \ \hat{\gamma}_C = c^T \hat{\gamma}_F \tag{12}$$

with c being a vector o regularization coefficients, satisfying

$$\frac{1}{n} \sum_j c_j^2 \to e < \infty \tag{13}$$

$$\sum_j (1 - c_j)^2 \Delta_j^2 \to 0 \ ; \ \sum_j (1 - c_j)^2 \gamma_j^2 \to 0 \tag{14}$$

When the set B is explicitly defined as an ellipsoid, such estimators can always be found by appropriately choosing c based on a parametric description of B (see [6]). However, this form of regularization is not related to the discriminant problem and a more natural choice seems to be to use some discriminant variable selection algorithm (i.e., chose c such that $c_j = 1\ ,\ j \leq m\ ;\ c_j = 0\ ,\ j > m$, for an m enforcing (13) and (14)) that does not depend on the explicit specification of B.

One possibility is to use a variant of forward variable selection adding variables, one by one, to previous sets, $S_0 = \emptyset, S_1, ..., S_{m-1}$, based on the additional discrimination criterion

$$X^{2(l)}(j) = \frac{\left(\hat{\Delta}_F(j) - \hat{\Sigma}(j, S_{l-1})\hat{\Sigma}^{-1}(S_{l-1}, S_{l-1})\hat{\Delta}_F(S_{l-1})\right)^2}{\hat{\Sigma}(j,j) - \hat{\Sigma}(j, S_{l-1})\hat{\Sigma}^{-1}(S_{l-1}, S_{l-1})\hat{\Sigma}(S_{l-1}, j)} \qquad (15)$$

Similarly to traditional forward selection, we stop adding variables when $max_{j \notin S_{l-1}} X^{2(l)}(j)/(1/n_0 + 1/n_1)$ falls below some quantil, $\chi^2_{1;1-\alpha}$, of a Qui-Square distribution with one degree of freedom. Furthermore, regardless of the value of the $X^{2(l)}(j)$ maximum, we always stop the selection process when $l = Mn$ for some, previously chosen, M constant.

The adjustment described above ensures that condition (13) is always satisfied, while conditions (14) are automatically satisfied (with probability one) for any such procedure with strictly positive α, since we are assuming that $\Delta \in l_2(N)$ and $n \to \infty$.

5 Finite Sample Performance

In order to evaluate the performance of δ_{Fct_q} in finite samples we performed a small simulation experiment with the following design.

We considered balanced samples with two combinations for the number of variables and sample size, $(p = 100, n = 200)$, and $(p = 100, n = 50)$. The first condition intends to illustrate a more traditional situation where the ratio n/p, although relatively small, is still larger than 1, while the second condition illustrates moderate dimensionality problems with $p > n$. For each combination of n and p we considered the following five data generating processes:

Condition A - All variables are independent.

Conditions B, C, D - Variables are generated according to model (5) with $q = 1$(Condition B), $q = 20$ (Condition C) and $q = p$ (Condition D).

Condition E - Variables are generated according to a factor model with p factors, with all specific variances set to 0.

In conditions B, C, D, and E, factor loadings were generated randomly according to an uniform U(0,1) distribution, and then normalized in order to achieve a pre-specified communality level. This level was set to 0.5 in conditions B, C and D, while in conditions A and E it was respectively equal to 0 and 1. In all conditions we assumed that 90 percent of the variables represented noise, having equal population means (set to 0) in both groups. For the remaining 10 percent (the signal), we set the means in the first group to 0, and in the second group to the geometic sequence $\mu_1 = (\nu, 0.9\ \nu, 0.9^2\ \nu, ...)$ where the constant ν was chosen in order to ensure a Mahalanobis distance between group centroids equal to 3. With this choice the expected rate of the theoretical Bayes rule is equal to $1 - \Phi(1.5) = 0.0668$.

We compared six different empirical classification rules, corresponding to two variants of the δ_F (or δ_{MF}), δ_I, and δ_{Fct_q} with $q = 1$, rules. In the first variant we used the $\hat{\Delta}_F$ and $\hat{\gamma}_F$ estimators for all rules (i.e., used all variables without any selection or shrinkage), while in the second variant we used the $\hat{\Delta}_C$ and $\hat{\gamma}_C$ estimators described in the previous section. In this later case we set α to 0.05 and assumed that the constant M was larger than 2 which, for the sample sizes considered, implied that our regularization reduced to traditional forward selection where in (15), $\hat{\Sigma}^{-1}$ was set to $\hat{\Sigma}_F^{-1}, \hat{\Sigma}_F^{-}, \hat{\Sigma}_I^{-1}$ or $\hat{\Sigma}_{Fctq}^{-1}$, respectively for the $\delta_F, \delta_{MF}, \delta_I$ and δ_{Fct_q} rules.

We then generated 100 independent training samples, used them to establish the empirical rules, and evaluated these rules on one, independently generated, balanced validation sample with 100 000 observations. The average misclassification rates in the validation sample are shown in Figures 1 and 2, and the K_0, K_{0F_q} constants and corresponding bounds on asymptotic error rates are presented in Table 1.

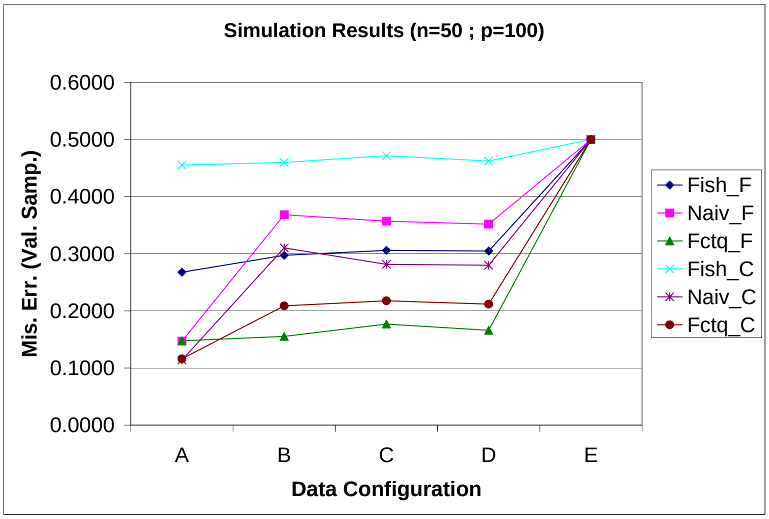

Fig. 1. Simulation Results – n=p/2

Table 1 illustrates some of the large differences that can occur between the K_0 and K_{0F_q} constants which here, with randomly generated factor loadings,

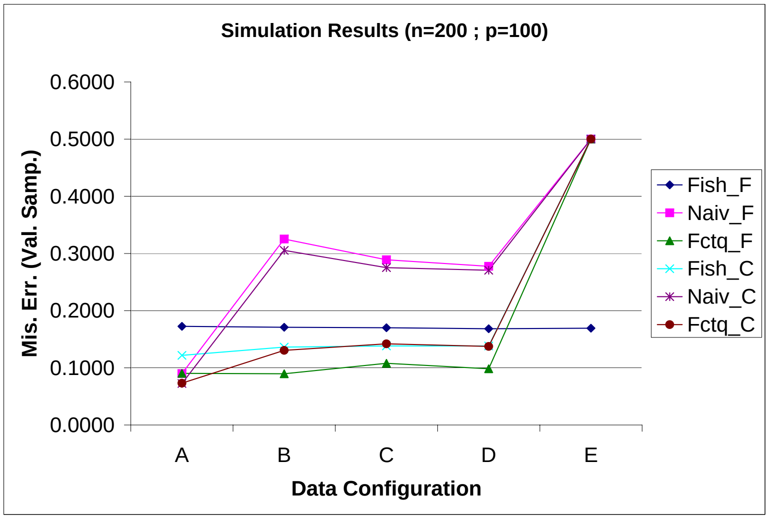

Fig. 2. Simulation Results – n=2p

Table 1. Theoretical constants and their asymptotic error rates

	Data Conditions				
	A	B	C	D	E
K_0	1	101.0	75.9	76.1	$6.57 * 10^8$
Naiv. As. Err.	0.067	0.384	0.367	0.367	0.500
K_{0F_q}	1	1	3.82	1.98	$8.14 * 10^6$
Fct_q As. Err.	0.067	0.067	0.112	0.078	0.499

lead in some cases to asymptotic bounds on error rates more than four times higher for the naive than for the δ_{fctq} rule. The actual error rates of the naive rules in the validation sample were often somehow smaller than predicted by their asymptotic bounds. This is not totally unexpected since the theoretical bounds reflect worst-case behavior. However, these bounds seemed to be good indicators of the order of magnitude of the true error rates.

We can see in Figure 1 that for data condition E with $n = p/2$, none of methods had a misclassification error meaningfully below than 0.5. The Fisher rule using all 100 variables was reasonably effective for this condition when $n = 2p$ (see Figure 2), but none of the other methods was able to benefit from the larger sample size. Notably, for this condition the variable selection procedure was not helpful at all, and all rules using the $\hat{\Delta}_C$ and $\hat{\gamma}_C$ estimators had misclassification rates close to fifty percent. This is a condition that was chosen exactly because of its inherent difficultly, which was confirmed by this results. However, we conjecture that similar structures might not be common in real life applications, where we expect most variables to have some specific variability and a dependence structure that may be explained by a lower-dimensional set of common factors.

On the other extreme, we see that both the δ_I and δ_{fctq} rules using the $\hat{\Delta}_C$ and $\hat{\gamma}_C$ estimators were quite effective when the data was indeed independent (data condition A), with error rates coming close to the one of the optimal rule when $n = 2p$. However, their performance worsened when the data was generated by a factor model with correlated variables and positive specific variances (data conditions B, C and D), where they were both outperformed by the δ_{fctq} rule using all 100 variables.

This result, in conjunction with those previously reported for condition E, suggests that, for these dimensionalities and sample sizes, any improvement due to the noise elimination in the $\hat{\Delta}_C$, $\hat{\gamma}_C$ estimators is often outweighted by the difficulty in separating discriminating variables from noisy ones. Curiously, the naive rule did not seem to be strongly affected by this problem, with the variant using $\hat{\Delta}_C$ and $\hat{\gamma}_C$ always giving better results than the one based on $\hat{\Delta}_F$ and $\hat{\gamma}_F$, with clear differences when $n = p/2$. However, this apparent better capacity of the δ_I rule in identifying important variables, was not enough to compensate for the consequences of ignoring the true correlations generated in conditions B, C and D. In higher dimensionalities, where it is likely that the importance of noise elimination will increase, some of these conclusions might be reversed. We leave a thorough investigation of this issue to future research.

The most interesting results are those concerning data conditions C and D. In these conditions, that we believe to be the more realistic ones, each variable has a variability explained in part by a common factor structure and in part by its own characteristics. In both conditions the true intrinsic dimensionality of the underlying model is considerably higher than the one assumed by the δ_{fctq} rule, although in condition C (but not in D) is smaller than the total number of variables. In both cases the variant of the δ_{fctq} rule that considers all available variables gave the best results with an expected error rate that was close the corresponding rate for the conditions (A and B) where the assumed model coincided with the data generating process. We consider these results to be particularly encouraging and have as top research priority to investigate if they still hold for higher dimensionalities and real-world data sets.

6 Discussion and Perspectives for Further Research

We proposed a new class of linear discriminant rules capable of incorporating information regarding correlation structures in problems with more variables than observations. Its main distinctive feature is the use of covariance estimates derived from low-dimensional factor models. Asymptotic properties and moderate dimensionality simulation results suggest that these rules can be quite effective under data conditions where Fisher's and naive discrimination rules perform poorly.

The asymptotic properties of the new rules require the use of shruken or truncated estimators of mean differences. However, our simulation results suggest that in problems with moderate dimensionalities the use of such regularized mean estimators might be counterproductive. For larger dimensionalities, where the damming effects of noisy variables are bound to be more serious, some fine tuning of regularization schemes may be required. This is an issue that we will address in future research.

In the present from, the proposed rules can be computationally too demanding for very high dimensional problems (with several thousand variables) common in genetic and microarray applications. Variants that try to alliviate the computational burden, while retaining some of its desirable statistical properties, are currently under investigation.

Other avenues of future research include the evaluation of the proposed rules in real-world data sets, and the development of generalizations to problems with more than two groups and to quadratic heterocedastic discrimination problems.

A Proof of Theorem 1

In this Appendix we demonstrate the claim we've made in Theorem 1, that we repeat here for convenience

$$lim\ sup_{n\to\infty}\ \overline{W}_{\Gamma_{Fct_q}}(\delta_{Fct_q}) \leq 1 - \Phi\left(\frac{\sqrt{K_{0F_q}}}{1+K_{0F_q}}\ c\right) \tag{16}$$

Firstly, we will introduce some notation.

Let $\delta^{\theta}_{Fct_q}$ be a population version of rule δ_{Fct_q} with $\hat{\theta} = (\hat{\Delta}_C, \hat{\gamma}_C, \hat{\Sigma}_{Fct_q})$ replaced by $(\Delta, \gamma, \Sigma_{Fct_q})$, and $\psi_{\Sigma}(\tilde{\Delta}, \tilde{\Sigma})$ denote the ratio

$$\psi_{\Sigma}(\tilde{\Delta}, \tilde{\Sigma}) = \frac{\tilde{\Delta}^T \tilde{\Sigma}^{-1} \tilde{\Delta}}{2(\tilde{\Delta}^T \tilde{\Sigma}^{-1} \Sigma \tilde{\Sigma}^{-1} \tilde{\Delta})^{1/2}} \tag{17}$$

where, depending on the rule being considered, $\tilde{\Delta}, \tilde{\Sigma}$, may represent parameters or estimators.

It can be easily shown that, for any linear rule of the form

$$Y = \delta_A(X, \tilde{\Delta}, \tilde{\Sigma}) = \mathbf{1}(\tilde{\Delta}^T \ \tilde{\Sigma}^{-1}\tilde{\gamma} > 0)$$

the posterior misclassification error probability is given by

$$W(\delta_A, \theta) = P_\theta(\delta_A(X, \tilde{\Delta}, \tilde{\Sigma}) = 1|\tilde{\Delta}, \tilde{\Sigma}, Y = 0) = 1 - \Phi(\psi_\Sigma(\tilde{\Delta}, \tilde{\Sigma})) \qquad (18)$$

In order to prove (16) we just need to show the following

$$\overline{W}_{\Gamma_{Fct_q}}(\delta^\theta_{Fct_q}) \le 1 - \Phi\left(\frac{\sqrt{K_{0F_q}}}{1 + K_{0F_q}}\ c\right) \qquad (19)$$

$$max_{\Gamma_{Fct_q}}\left\{\left|\psi_\Sigma(\hat{\Delta}_C, \hat{\Sigma}_{Fct_q}) - \psi_\Sigma(\Delta, \Sigma_{Fct_q})\right|\right\} \xrightarrow{P} 0 \qquad (20)$$

$$max_{\Gamma_{Fct_q}}\left\{\left|\psi_\Sigma(\hat{\Delta}, \hat{\Sigma}_{Fct_q}) - \psi_\Sigma(\Delta, \hat{\Sigma}_{Fct_q})\right|\right\} \xrightarrow{P} 0 \qquad (21)$$

where the convergence in probability of (20) and (21) is uniform over Γ_{Fct_q}.

Conditions (19), (20) and (21) imply (16) because, exactly by the same arguments as made by Bickel and Levina in the demosntration of their theorem 1, part b (see [1], pp. 995-996), (20) and (21) are enough to establish convergence of $\overline{W}_{\Gamma_{Fct_q}}(\delta_{Fct_q})$ to $\overline{W}_{\Gamma_{Fct_q}}(\delta^\theta_{Fct_q})$.

We now proceed to the demonstration of (19), following again the same line of reasoning as in [1].

We first note that

$$\overline{W}_{\Gamma_{Fct_q}}(\delta^\theta_{Fct_q}) = max_{\Gamma_{Fct_q}}(1 - \Phi(\psi_\Sigma(\Delta, \Sigma_{Fct_q}))\ = \ 1 - \Phi(min_{\Gamma_{Fct_q}}\ \psi_\Sigma(\Delta, \Sigma_{Fct_q})\ )$$

Next, we write $min_{\Gamma_{Fct_q}}\ \psi_\Sigma(\Delta, \Sigma_{Fct_q})$ as

$$min_{\Gamma_{Fct_q}}\ \psi_\Sigma(\Delta, \Sigma_{Fct_q}) = \frac{c}{2} min_{\{\Gamma_{Fct_q}\ :\ \Delta^T\Sigma^{-1}\Delta = c^2\}} \frac{\psi_\Sigma(\Delta, \Sigma_{Fct_q})}{\psi_\Sigma(\Delta, \Sigma)} = \frac{c}{2} min_{\{\Gamma_{Fct_q}\ :\ \Delta^T\Sigma^{-1}\Delta = c^2\}} \frac{\Delta^T(\Sigma^{-1/2}_{Fct_q})^T\Sigma^{-1/2}_{Fct_q}\Delta}{[(\Delta^T(\Sigma^{-1/2}_{Fct_q})^T\Sigma_{0Fct_q}\Sigma^{-1/2}_{Fct_q}\Delta)(\Delta^T(\Sigma^{-1/2}_{Fct_q})^T\Sigma^{-1}_{0Fct_q}\Sigma^{-1/2}_{Fct_q}\Delta)]^{1/2}}$$

which is true because $\psi_\Sigma(\Delta, \Sigma_{Fct_q})$ always reaches its minimum on Γ_{Fct_q} at some θ such that $\Delta^T\Sigma^{-1}\Delta = c^2$, and $\psi_\Sigma(\Delta, \Sigma) = (1/2)(\Delta^T\Sigma^{-1}\Delta)^{1/2}$.

Therefore, defining $\Delta_{0Fct_q} = \Sigma^{-1/2}_{Fct_q}\Delta$, it follows that

$$min_{\Gamma_{Fct_q}} \ \psi_{\Sigma}(\Delta, \Sigma_{Fct_q}) =$$

$$\frac{c}{2} min_{\{\Gamma_{Fct_q} \ : \ \Delta^T \Sigma^{-1} \Delta = c^2\}} \frac{\Delta^T_{0Fct_q} \Delta_{0Fct_q}}{[(\Delta^T_{0Fct_q} \Sigma_{0Fct_q} \Delta_{0Fct_q})(\Delta^T_{0Fct_q} \Sigma^{-1}_{0Fct_q} \Delta_{0Fct_q})]^{1/2}} \geq$$

$$\frac{c}{2} \ \frac{2\left[\lambda_{max}(\Sigma_{0Fct_q})\lambda_{min}(\Sigma_{0Fct_q})\right]^{1/2}}{\lambda_{min}(\Sigma_{0Fct_q}) + \lambda_{max}(\Sigma_{0Fct_q})} \ = \ \frac{\sqrt{K_{0F_q}}}{1 + K_{0F_q}} \ c$$

where the inequality above follows from Kantorovich inequality, which states that for any vector, v, and positive defenite matrix, M, of conformable dimensions

$$\frac{(v^T v)^2}{(v^T M v)(v^T M^{-1} v)} \ \geq \frac{4 \ \lambda_{min}(M) \ \lambda_{max}(M)}{[\lambda_{min}(M) + \lambda_{max}(M)]^2}$$

This establishes (19), so now we turn our attention to the demonstration of (20) and (21).

We first claim that when

$$max_{\Gamma_{Fct_q}} max_{j,l} |\Sigma^{-1}_{Fct_q}(j,l) - \hat{\Sigma}^{-1}_{Fct_q}(j,l)| \xrightarrow{P} 0. \qquad (22)$$

hold, then (20) and (21) are true.

To verify this claim, expand $1/p^4$ times the numerator of $\psi_{\Sigma}(\tilde{\Delta}, \tilde{\Sigma})$ around $\hat{\Delta}_C$ and $\hat{\Sigma}^{-1}_{Fctq}$, as

$$\frac{1}{p^4}\left(\hat{\Delta}_C + \epsilon_1 \ e\right)^T \left(\hat{\Sigma}^{-1}_{Fctq} + \epsilon_2 \ E\right)\left(\hat{\Delta}_C + \epsilon_1 \ e\right) = \frac{1}{p^4}(\hat{\Delta}^T_C \ \hat{\Sigma}^{-1}_{Fctq} \ \hat{\Delta}_C +$$
$$2\epsilon_1 \ e^T \hat{\Sigma}^{-1}_{Fctq} \ \hat{\Delta}_C + \ \epsilon_2 \ \hat{\Delta}^T_C \ E \ \hat{\Delta}_C \ + \ 2\epsilon_1\epsilon_2 \ e^T E \ \hat{\Delta}_C \ + \ \epsilon_1^2 \epsilon_2 \ e^T E \ e\Big) =$$
$$\frac{1}{p^4}\left(\hat{\Delta}^T_C \ \hat{\Sigma}^{-1}_{Fctq} \ \hat{\Delta}_C\right) + O(\epsilon_1 + \epsilon_2) \qquad (23)$$

and $1/p^4$ times the denominator of $\psi_{\Sigma}(\tilde{\Delta}, \tilde{\Sigma})$ as

$$\frac{2}{p^4}\left[\left(\hat{\Delta}_C + \epsilon_1 \ e\right)^T \left(\hat{\Sigma}^{-1}_{Fctq} + \epsilon_2 \ E\right) \Sigma \left(\hat{\Sigma}^{-1}_{Fctq} + \epsilon_2 \ E\right)\left(\hat{\Delta}_C + \epsilon_1 \ e\right)\right]^{1/2} =$$
$$\frac{2}{p^4}\Big[(\hat{\Delta}^T_C \ \hat{\Sigma}^{-1}_{Fctq} \ \Sigma \ \hat{\Sigma}^{-1}_{Fctq} \ \hat{\Delta}_C \ + \ 2\epsilon_1 \ e^T \hat{\Sigma}^{-1}_{Fctq} \ \Sigma \ \hat{\Sigma}^{-1}_{Fctq} \ \hat{\Delta}_C \ +$$
$$2\epsilon_2 \ \hat{\Delta}^T_C \ E \ \Sigma \ \hat{\Sigma}^{-1}_{Fctq} \ \hat{\Delta}_C \ + \ \epsilon_1^2 \ e^T \hat{\Sigma}^{-1}_{Fctq} \ \Sigma \ \hat{\Sigma}^{-1}_{Fctq} \ e \ +$$
$$\epsilon_2^2 \ \hat{\Delta}^T_C \ E \ \Sigma \ E \ \hat{\Delta}_C \ + \ 4\epsilon_1\epsilon_2 \ e^T E \ \Sigma \ \hat{\Sigma}^{-1}_{Fctq} \ \hat{\Delta}_C \ +$$
$$2\epsilon_1^2\epsilon_2 \ e^T E \ \Sigma \ \hat{\Sigma}^{-1}_{Fctq} \ e \ + \ 2\epsilon_1\epsilon_2^2 \ e^T E \ \Sigma \ E \ \hat{\Delta}_C \ + \ \epsilon_1^2\epsilon_2^2 \ e^T E \ \Sigma \ E \ e \ )\Big]^{1/2} =$$
$$\frac{2}{p^4}\left(\hat{\Delta}^T_C \ \hat{\Sigma}^{-1}_{Fctq} \ \Sigma \ \hat{\Sigma}^{-1}_{Fctq} \ \hat{\Delta}_C \ \right)^{1/2} + \ O(\epsilon_1 + \epsilon_2) \qquad (24)$$

where e and E are the unitary vector an matrix, and in the last equalities of (23) and (24) we used the facts that $\hat{\Delta}_C^T\ \hat{\Sigma}_{Fctq}^{-1}\ \Sigma\ \hat{\Sigma}_{Fctq}^{-1}\ \hat{\Delta}_C$ is bounded away from zero (see inequality (4.13) in [1]), and that the maximum absolute-value elements of $\hat{\Delta}_C$, Σ, $\Sigma\hat{\Sigma}_{Fctq}^{-1}$ and $\hat{\Sigma}_{Fctq}^{-1}\Sigma\hat{\Sigma}_{Fctq}^{-1}$ are all bounded by a constant that does not depend on p. This last claim is a consequence of $\hat{\Delta}_C \in l_2(N)$ and known properties of symmetric matrix norms, namely $max_{jl}|M(j,l)| \leq ||M||_2 = \sqrt{\lambda_{max}(M)}$, $||M\ N||_2 \leq ||M||_2||N||_2$, and the conditions $\lambda_{max}(\Sigma) \leq k_2$ and $\lambda_{max}(\hat{\Sigma}_{Fctq}^{-1}) = \lambda_{min}^{-1}(\hat{\Sigma}_{Fctq}) \leq k_0^{-1}$.

From these expansions, and the property $\hat{\Delta}_C^T\ \hat{\Sigma}_{Fctq}^{-1}\ \Sigma\ \hat{\Sigma}_{Fctq}^{-1}\ \hat{\Delta}_C > h$ with strictly positive h, it follows that

$$\Psi_\Sigma\left(\hat{\Delta}_C + \epsilon_1 e, \hat{\Sigma}_{Fctq}^{-1} + \epsilon_2 E\right) \ = \ \Psi_\Sigma\left(\hat{\Delta}_C, \hat{\Sigma}_{Fctq}^{-1}\right) + \ O(\epsilon_1 + \epsilon_2) \tag{25}$$

which, together with the condition $E_\theta||\hat{\Delta}_C - \Delta||^2 = o(1)$, is sufficient to show that (22) implies (20) and (21).

In order to prove (22) we first remark that, as $\hat{\Sigma}_F$ follows the Wishart distribution, $W(n-2,\Sigma)$, (see e.g. [8]), $\hat{\Sigma}_F(j,l)$ can always be written as

$$\hat{\Sigma}_F(j,l) = \sqrt{\Sigma(j,j)\Sigma(l,l)}\ \sum_{i=1}^{n-2} Z_{ij}Z_{il} \tag{26}$$

where (Z_{ij}, Z_{il}) are independent bivariate normal vectors with unit variances and correlation $\rho(j,l) = \Sigma(j,l)/\sqrt{\Sigma(j,j)\Sigma(l,l)}$.

Then, by Lemma 4 in [1], it follows that

$$\begin{aligned} &P_\theta\left(max_{j,l}\left|\hat{\Sigma}_F(j,l) - \Sigma(j,l)\right| \geq \epsilon\right) = \\ &P_\theta\left(max_{j,l}\left|\frac{\hat{\Sigma}_F(j,l)}{\sqrt{\Sigma(j,j)\Sigma(l,l)}} - \rho(j,l)\right| \geq \epsilon\right) \ \leq \ \frac{p(p+1)}{2}\ e^{-(n-2)c(\epsilon)} \end{aligned} \tag{27}$$

for a known positive constant, $c(\epsilon)$, not depending on n or p.

Therefore, it follows that when $(ln\ p^2)/n \to 0$

$$max_{j,l}|\hat{\Sigma}_F(j,l) - \Sigma(j,l)| \xrightarrow{P} 0 \tag{28}$$

To show that a similar property holds for $\hat{\Sigma}_{Fct_q}$ we make a Taylor series expansion of $\beta(j,a)$ and $\Omega(j,j)$ as

$$\begin{aligned} \hat{\beta}(j,a) = \beta(j,a) + \sum_{j',l'} \frac{\partial\beta(j,a)}{\partial\Sigma(j',l')}\left(\hat{\Sigma}_F(j',l') - \Sigma(j',l')\right) + \\ o\left(max_{j',l'}|\hat{\Sigma}_F(j',l') - \Sigma(j',l')|\right) \end{aligned} \tag{29}$$

$$\hat{\Omega}(j,j) = \Omega(j,j) + \sum_{j',l'} \frac{\partial \Omega(j,j)}{\partial \Sigma(j',l')} \left(\hat{\Sigma}_F(j',l') - \Sigma(j',l')\right) + o\left(max_{j',l'}|\hat{\Sigma}_F(j',l') - \Sigma(j',l')|\right) \quad (30)$$

which, in view of (28) and conditions (10), imply that with probability tending to one

$$|\hat{\beta}(j,a) - \beta(j,a)| \leq M \ max_{j',l'}|\hat{\Sigma}_F(j',l') - \Sigma(j',l')| \quad (31)$$

$$|\hat{\Omega}(j,j) - \Omega(j,j)| \leq M \ max_{j',l'}|\hat{\Sigma}_F(j',l') - \Sigma(j',l')| \quad (32)$$

for some common finite $M > max_{j,a}\left(\sum_{j',l'}|\frac{\partial\beta(j,a)}{\partial\Sigma(j',l')}|, \sum_{j',l'}|\frac{\partial\Omega(j,j)}{\partial\Sigma(j',l')}|\right)$.

Furthermore, since $\hat{\Sigma}_{Fct_q}(j,j) = \hat{\Omega}^2(j,j) + \sum_{a=1}^{q}\hat{\beta}^2(j,a)$, $\hat{\Sigma}_{Fct_q}(j,l) = \sum_{a=1}^{q}\hat{\beta}(j,a)\hat{\beta}(l,a)$ for $(j \neq l)$, and q is fixed, it follows that

$$max_{j,l}|\hat{\Sigma}_{Fct_q}(j,l) - \Sigma_{Fct_q}(j,l)| \xrightarrow{P} 0 \quad (33)$$

All that it's left, is to show that convergence of $max_{j,l}\hat{\Sigma}_{Fct_q}(j,l)$ implies convergence of $max_{j,l}\hat{\Sigma}^{-1}_{Fct_q}(j,l)$. That will follow from

$$max_{j,l}|\hat{\Sigma}^{-1}_{Fct_q}(j,l) - \Sigma^{-1}_{Fct_q}(j,l)| \leq C \ max_{j,l}|\hat{\Sigma}_{Fct_q}(j,l) - \Sigma_{Fct_q}(j,l)| \quad (34)$$

for some positive constant, C, not depending on n or p.

To show that (34) is indeed true, we note that from (6)

$$\hat{\Sigma}^{-1}_{Fct_q}(j,j) = \hat{\Omega}^{-2}(j,j) - \sum_{a,b=1}^{q}\hat{N}^{-1}(a,b)\hat{M}^2(j,a) \quad (35)$$

$$\hat{\Sigma}^{-1}_{Fct_q}(j,l) = \sum_{a,b=1}^{q}\hat{N}^{-1}(a,b)\hat{M}(j,a)\hat{M}(l,a) \ (l \neq j) \quad (36)$$

where $\hat{M} = \hat{\Omega}^{-1}\hat{\beta}$ and $\hat{N} = I_q + \hat{M}^T\hat{M}$.

But since q is fixed and by assumption $\hat{\Omega}(j,j) > k_0$, we just need to show that for the q-dimensional matrix N

$$max_{a,b}|\hat{N}^{-1}(a,b) - N^{-1}(a,b)| \leq C \ max_{a,b}|\hat{N}(a,b) - N(a,b)| \quad (37)$$

Inequality (37) follows from known results in the analysis of matrix perturbations, in particular from (see theorem 2.3.4 in [5])

$$max_{a,b}|\hat{N}^{-1}(a,b) - N^{-1}(a,b)| \;\leq\; ||\hat{N}^{-1} - N^{-1}||_2 \;\leq\; \frac{||\hat{N}^{-1}||_2^2}{1-r}||\hat{N} - N||_2 \;=\; \frac{\lambda_{min}^{-1}(\hat{N})}{1-r}||\hat{N} - N||_2 \;\leq\; \frac{1}{1-r}||\hat{N} - N||_2 \;\leq\; \frac{q}{1-r}max_{a,b}|\hat{N}(a,b) - N(a,b)|$$

where $r = ||\hat{N}^{-1}(N - \hat{N})||_2 = ||\hat{N}^{-1}N - I_q||_2 < 1$ for $\hat{N}$ close enough to N, and we used the well known inequality (for any A, a q-dimensional square matrix) $||A||_2 \leq q \; max_{a,b}A(a,b)$, and the fact that for any matrix of the form $A = I_q + B^TB$, it is always true that $\lambda_{min}(A) \geq 1$.

This proves (34) and completes the demonstration of the Theorem.

References

1. P.J. Bickel and E. Levina. Some Theory for Fisher's Linear Discriminant Function, "Naive Bayes" and some Alternatives when there are many more Variables than Observations. *Bernoulli* 10, 6: 989-1010, 2004.
2. P. Domingos and M. Pazzani. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss. *Machine Learning* 29: 103-130, 1997.
3. S. Dudoit, J. Fridlyand and T.P. Speed. Comparison of Discriminant Methods for the Classification of Tumors using Gene Expression Data. *Journal of the American Statistical Association* 97: 77-87, 2002.
4. J. Fan and Y. Fan. High Dimensional Classification using Features Annealed Independence Rules. *Annals of Statistics* 38, 6: 2605-2637, 2008.
5. G.H. Golub, and C.F. Van Loan. *Matrix Computations.* Johns Hopkins, Baltimore, 1996.
6. I.M. Johnstone. Function Estimation and Gaussian Sequence Models. *Unpublished monograph*, http://www-stat.stanford.edu/~imj, 2002.
7. G.J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition.* Wiley, New York, 1992.
8. G.A.F. Seber. *Multivariate Observations.* Wiley, New York, 1984.