Bernoulli **10**(6), 2004, 989–1010

Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations

PETER J. BICKEL[1] and ELIZAVETA LEVINA[2]

[1]*Department of Statistics, University of California, Berkeley CA 94720-3860, USA.*
E-mail: bickel@stat.berkeley.edu
[2]*Department of Statistics, University of Michigan, Ann Arbor MI 48109-1092, USA.*
E-mail: elevina@umich.edu

We show that the 'naive Bayes' classifier which assumes independent covariates greatly outperforms the Fisher linear discriminant rule under broad conditions when the number of variables grows faster than the number of observations, in the classical problem of discriminating between two normal populations. We also introduce a class of rules spanning the range between independence and arbitrary dependence. These rules are shown to achieve Bayes consistency for the Gaussian 'coloured noise' model and to adapt to a spectrum of convergence rates, which we conjecture to be minimax.

Keywords: Fisher's linear discriminant; Gaussian coloured noise; minimax regret; naive Bayes

1. Introduction

It has long been appreciated in machine learning practice (see, for example Lewis 1998; Domingos and Pazzani 1997) that in classification problems in which the number of covariates is large, rules which use the evidently invalid assumption that the covariates are independent often perform better than rules which try to estimate dependence between covariates in the construction of the classifier. We were struck by this phenomenon in some problems of texture classification (Levina, 2002), though, unfortunately, the context we were working in was far too complicated for direct analysis. The same phenomenon has been reported for microarray data (Dudoit *et al.*, 2002), where ignoring correlations between genes led to better classification results.

To study this practical success analytically, we decided to explore the power of two classical classifiers, the Fisher linear discriminant function and the so-called 'naive Bayes' rule, which assumes independence in the simple context of the multivariate Gaussian model. To understand what happens qualitatively, we let, in our asymptotics, both the dimension p of the vector observations and the size of the training sample n to be large, with p quite possibly much larger than n. We present our approach and results in Section 2. Our results are of two types. In Section 2 we show that, on the basis of a worst-case analysis, for large p, naive Bayes can indeed greatly outperform the linear discriminant function. Section 3 points out the connection between the conditions that guarantee results of Section 2 and the spectral density. The surprisingly good performance of naive Bayes led us to consider a spectrum of rules spanning the range between assuming full independence and arbitrary dependence. We present these rules in Section 4, where we also formulate the Bayes consistency and minimax regret problems in the context of the coloured Gaussian noise model. We show that, using modifications to our rules, we can adapt to a spectrum of rates which we conjecture to be minimax. We conclude in Section 5 with a discussion of the relation of our work to that of Greenshtein and Ritov (2004), and more generally, to the criterion of 'sparsity' of the parameters – see Donoho *et al.* (1995). Details of the proofs of necessary lemmas are given in Section 6.

2. Model and first results

Consider the problem of discriminating between two classes with p-variate normal distributions $N_p(\boldsymbol{\mu}_0, \Sigma)$ and $N_p(\boldsymbol{\mu}_1, \Sigma)$. A new observation $\mathbf{X}$ is to be assigned to one of these two classes. If $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$ and Σ are known then the optimal classifier is the Bayes rule:

$$\delta(\mathbf{X}) = \mathbf{1}\left\{\log\frac{f_1(\mathbf{X})}{f_0(\mathbf{X})} > 0\right\} = \mathbf{1}\{\boldsymbol{\Delta}^{\mathrm{T}}\Sigma^{-1}(\mathbf{X} - \boldsymbol{\mu}_\cdot) > 0\}, \tag{2.1}$$

where the class prior probabilities are assumed equal, f_0 and f_1 are the densities of $N_p(\boldsymbol{\mu}_0, \Sigma)$ and $N_p(\boldsymbol{\mu}_1, \Sigma)$, respectively, and

$$\boldsymbol{\Delta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0, \qquad \boldsymbol{\mu}_\cdot = \frac{1}{2}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1).$$

If we have independent observations from the two classes $\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in} (i = 0, 1)$, and estimators $\hat{\boldsymbol{\mu}}_0$, $\hat{\boldsymbol{\mu}}_1$ of the population means, then the quantities in (2.1) can be estimated by $\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0$, $\hat{\boldsymbol{\mu}}_\cdot = \frac{1}{2}(\hat{\boldsymbol{\mu}}_1 + \hat{\boldsymbol{\mu}}_0)$. Σ is estimated by the pooled estimate where the centring is at the classical $\hat{\boldsymbol{\mu}}_i = \overline{\mathbf{X}}_i = \frac{1}{n}\sum_{k=1}^{n}\mathbf{X}_{ik}$ (for $i = 0, 1$),

$$\hat{\Sigma} = \frac{1}{2(n-1)}\sum_{i=0}^{1}\sum_{k=1}^{n}(\mathbf{X}_{ik} - \overline{\mathbf{X}}_i)^{\mathrm{T}}(\mathbf{X}_{ik} - \overline{\mathbf{X}}_i).$$

Even though we assume equal sample sizes for convenience, all the results below extend trivially to unequal sample sizes n_0 and n_1 as long as $n_0 \to \infty$, $n_1 \to \infty$, and $n_0/(n_1 + n_0) \to \pi$, $0 < \pi < 1$. If we naturally assume that π is the probability of a new observation belonging to class 0, the rule is modified by replacing 0 in the indicator by $\log(n_0/n_1)$.

By convention, we always view $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$ as points in l_2 by adding 0s at the end.

Plugging all the parameter estimates directly into the Bayes rule (2.1) leads to the Fisher rule (FR),

$$\delta_{\mathrm{F}}(\mathbf{X}) = \mathbf{1}\{\hat{\boldsymbol{\Delta}}^{\mathrm{T}}\hat{\Sigma}^{-1}(\mathbf{X} - \hat{\boldsymbol{\mu}}_\cdot) > 0\}.$$

Alternatively, assuming independence of components and replacing off-diagonal elements of $\hat{\Sigma}$ with zeros leads to a new covariance matrix estimate,

$$\hat{D} = \mathrm{diag}(\hat{\Sigma}),$$

and a different discrimination rule, the independence rule (IR),

$$\delta_{\mathrm{I}}(\mathbf{X}) = \mathbf{1}\{\hat{\boldsymbol{\Delta}}^{\mathrm{T}}\hat{D}^{-1}(\mathbf{X} - \hat{\boldsymbol{\mu}}.) > 0\},$$

which is also known as naive Bayes. The first goal of this paper is to compare the performance of these two rules as $p \to \infty$, $n \to \infty$, and $p/n \to \gamma$ with $0 \leqslant \gamma \leqslant \infty$. We will compare the rules in terms of their worst-case performance. Let

$$\Gamma(c, k, B) = \{(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \Sigma) : \boldsymbol{\Delta}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\Delta} \geqslant c^2, k_1 \leqslant \lambda_{\min}(\Sigma) \leqslant \lambda_{\max}(\Sigma) \leqslant k_2,$$
$$\boldsymbol{\mu}_i \in B, i = 0, 1\},$$

where c, k_1, and k_2 are positive constants, $\lambda_{\min}(\Sigma)$, $\lambda_{\max}(\Sigma)$ are, respectively, the smallest and the largest eigenvalues of Σ, and B is the compact subset of l_2 given by

$$B = B_{\mathbf{a},d} = \left\{\boldsymbol{\mu} \in l_2 : \sum_{j=1}^{\infty} a_j\mu_j^2 \leqslant d^2\right\}.$$

Here, $a_j \to \infty$ and $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots)$. It is well known that B is a compact and that (see Pinsker's theorem in Johnstone 2002) if Σ is the identity, then for suitable r_{jn}, depending only on $\{a_j\}$, the jth component of $\boldsymbol{\mu}_i$ can be estimated by

$$\hat{\mu}_{ij} = (1 - r_{jn})_+\overline{X}_{ij}, \qquad i = 0, 1, \tag{2.2}$$

and

$$\max_{\Gamma} \mathrm{E}_{\boldsymbol{\theta}}\|\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i\|^2 = o(1). \tag{2.3}$$

The condition on eigenvalues guarantees that

$$\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} \leqslant K = \frac{k_2}{k_1}.$$

Then both Σ and Σ^{-1} are not ill-conditioned. The condition $\boldsymbol{\Delta}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\Delta} \geqslant c^2$ guarantees that the Mahalanobis distance between the two populations is at least c, so that c is a measure of difficulty of the classification problem. Let $\boldsymbol{\theta} = (\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \Sigma)$. Assume henceforth that $\mathbf{X} \sim N(\boldsymbol{\mu}_0, \Sigma)$. The symmetry of our rules makes the posterior probability of misclassification if the mean of $\mathbf{X}$ is $\boldsymbol{\mu}_0$ the same as that under $\boldsymbol{\mu}_1$.

For a rule δ and $\mathbf{X} \sim N(\boldsymbol{\mu}_0, \Sigma)$, define posterior error by

$$W(\delta, \boldsymbol{\theta}) = P_{\boldsymbol{\theta}}[\delta(\mathbf{X}) = 1|\mathbf{X}_{ik}, i = 0, 1, k = 1 \ldots n]$$

and the worst-case posterior error by

$$W_{\Gamma}(\delta) = \max_{\Gamma} W(\delta, \boldsymbol{\theta}).$$

Further, let

$$\overline{W}(\delta, \boldsymbol{\theta}) = P_\theta[\delta(\mathbf{X}) = 1]$$

be the misclassification error of δ, and

$$\overline{W}_\Gamma(\delta) = \max_\Gamma \overline{W}(\delta, \boldsymbol{\theta})$$

be the worst-case error. For the two rules δ_F and δ_I, the posterior errors can easily be computed as

$$W(\delta_\mathrm{F}, \boldsymbol{\theta}) = \overline{\Phi}(\Psi_\Sigma(\hat{\boldsymbol{\Delta}}, \hat{\Sigma})),$$

$$W(\delta_\mathrm{I}, \boldsymbol{\theta}) = \overline{\Phi}(\Psi_\Sigma(\hat{\boldsymbol{\Delta}}, \hat{D})),$$

where $\overline{\Phi} = 1 - \Phi$, Φ is the Gaussian cumulative distribution function, and

$$\Psi_\Sigma(\boldsymbol{\Delta}, M) = \frac{\boldsymbol{\Delta}^\mathrm{T} M^{-1}\boldsymbol{\Delta}}{2(\boldsymbol{\Delta}^\mathrm{T} M^{-1}\Sigma M^{-1}\boldsymbol{\Delta})^{1/2}}. \tag{2.4}$$

The behaviour of these errors is complex and has been studied extensively for the case of fixed p (for a review, see McLachlan 1992). It is well known that the rule δ_F is asymptotically optimal for this problem, that is,

$$\overline{W}_\Gamma(\delta_\mathrm{F}) \to \overline{\Phi}(c/2),$$

which is the Bayes risk, while $\overline{W}_\Gamma(\delta_\mathrm{I})$ converges to something strictly greater than the Bayes risk. If $p > n$, δ_F is not well defined since $\hat{\Sigma}^{-1}$ is not. We replace $\hat{\Sigma}^{-1}$ by $\hat{\Sigma}^{-}$, the Moore–Penrose inverse, obtained by finding the subspace of $\mathbb{R}^p$ spanned by the eigenvectors $\hat{\boldsymbol{\xi}}_1, \ldots, \hat{\boldsymbol{\xi}}_n$ corresponding to non-zero eigenvalues $\hat{\lambda}_1, \ldots, \hat{\lambda}_n$ of $\hat{\Sigma}$ and then defining

$$\hat{\Sigma}^{-} = n^{-1}\sum_{i=1}^{n} \frac{1}{\hat{\lambda}_i}\hat{\boldsymbol{\xi}}_i\hat{\boldsymbol{\xi}}_i^\mathrm{T}.$$

Let $\Sigma_0 = D^{-1/2}\Sigma D^{-1/2}$ be the correlation matrix of $\mathbf{X}$ and let

$$K_0 = \max_\Gamma \frac{\lambda_{\max}(\Sigma_0)}{\lambda_{\min}(\Sigma_0)}.$$

Note that $K_0 \leqslant K^2$ since $k_1 \leqslant \sigma_{ii} \leqslant k_2$ for all i.

Theorem 1. (a) *If* $p/n \to \infty$, *then* $\overline{W}_\Gamma(\delta_\mathrm{F}) \to \frac{1}{2}$.

(b) *If* $(\log p)/n \to 0$, *then*

$$\limsup_{n\to\infty} \overline{W}_\Gamma(\delta_\mathrm{I}) = \overline{\Phi}\left(\frac{\sqrt{K_0}}{1+K_0}c\right). \tag{2.5}$$

Note that if the matrix has eigenvalues going to 0 or ∞ as $p \to \infty$, then $K_0 \to \infty$, and $\limsup \overline{W}(\delta_\mathrm{I}) = \frac{1}{2}$, so the worst case of the rule is no better than random guessing. However, if Σ is a multiple of the identity so that $K_0 = 1$, then the bound gives the Bayes risk, as it should since in this case the IR is asymptotically optimal.

Remark. It is worth noting that even when $\boldsymbol{\Delta}$ and Σ are assumed known, the corresponding IR does not lose much in comparison to the Bayes rule. This remains true for the original IR under the conditions of Theorem 1 since then e_2 below is the limiting risk of IR. To see that, let

$$e_1 = \overline{\Phi}(\Psi_\Sigma(\boldsymbol{\Delta}, \Sigma)) = \overline{\Phi}\left(\frac{1}{2}(\boldsymbol{\Delta}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\Delta})^{1/2}\right),$$

$$e_2 = \overline{\Phi}(\Psi_\Sigma(\boldsymbol{\Delta}, D)) = \overline{\Phi}\left(\frac{1}{2}\frac{\boldsymbol{\Delta}^{\mathrm{T}}D^{-1}\boldsymbol{\Delta}}{(\boldsymbol{\Delta}^{\mathrm{T}}D^{-1}\Sigma D^{-1}\boldsymbol{\Delta})^{1/2}}\right)$$

be the errors of the two rules when $\boldsymbol{\Delta}, \Sigma$ and $D = \mathrm{diag}(\Sigma)$ are known. If we write $\boldsymbol{\Delta}_0 = D^{-1/2}\boldsymbol{\Delta}$, then the efficiency of the IR relative to the FR is determined by the ratio r of the arguments of $\overline{\Phi}$, where

$$r = \frac{\Psi_\Sigma(\boldsymbol{\Delta}, D)}{\Psi_\Sigma(\boldsymbol{\Delta}, \Sigma)} = \frac{(\boldsymbol{\Delta}_0^{\mathrm{T}}\boldsymbol{\Delta}_0)}{[(\boldsymbol{\Delta}_0^{\mathrm{T}}\Sigma_0\boldsymbol{\Delta}_0)(\boldsymbol{\Delta}_0^{\mathrm{T}}\Sigma_0^{-1}\boldsymbol{\Delta}_0)]^{1/2}}. \tag{2.6}$$

A bound on this quantity can be obtained from the Kantorovich inequality (quoted here from Luenberger 1984): let Q be any positive definite symmetric $p \times p$ matrix. Then, for any vector $\mathbf{v}$,

$$\frac{(\mathbf{v}^{\mathrm{T}}\mathbf{v})^2}{(\mathbf{v}^{\mathrm{T}}Q\mathbf{v})(\mathbf{v}^{\mathrm{T}}Q^{-1}\mathbf{v})} \geqslant \frac{4aA}{(a+A)^2}$$

where a is the smallest eigenvalue of Q, and A is the largest. Applying this inequality to (2.6), we obtain

$$r \geqslant \frac{2\sqrt{K_0}}{1+K_0} \tag{2.7}$$

and the error of the IR can be bounded by

$$e_1 \leqslant e_2 \leqslant \overline{\Phi}\left(\frac{2\sqrt{K_0}}{1+K_0}\overline{\Phi}^{-1}(e_1)\right).$$

The actual loss in efficiency is not very large: Figure 1 presents plots of the bound as a function of the Bayes risk e_1 for several values of K_0. For moderate K_0, one can see that the performance of the IR is comparable to that of the FR. Note that the bounds represent the worst-case performance, so the actual results may be and in fact should typically be better. In practice, K_0 cannot be estimated reliably from data, since the estimated pooled correlation matrix is only of rank $2(n-1)$. The range of non-zero eigenvalues of the estimated correlation matrix, however, does give one a rough idea about the value of K_0. For instance, in the leukaemia data set discussed in Dudoit *et al.* (2002), $K_0 \approx 30$, so one can expect the naive Bayes rule to perform reasonably well (and it does in fact perform much better than the Fisher rule).

Before proceeding to the proof of Theorem 1, we state a necessary lemma, whose proof for

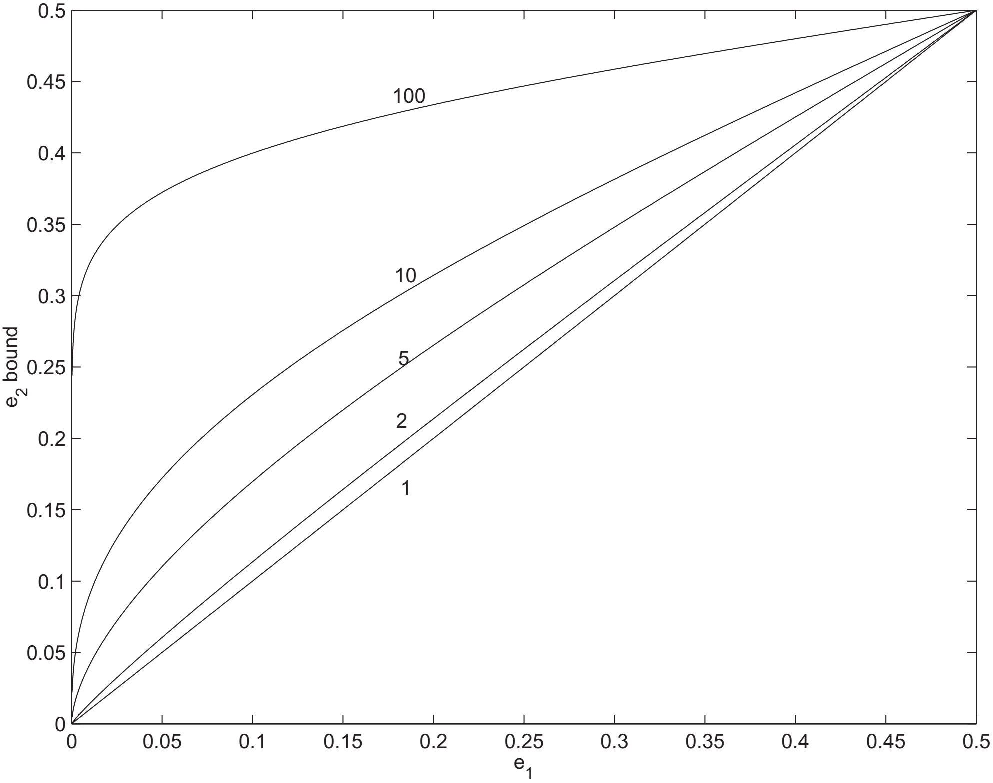

Figure 1. The bound on the risk of the IR as a function of the Bayes risk. The numbers over the curves show the value of K_0.

$\Sigma = I$ appears, for instance, in Johnstone (2002). We establish this extension in Section 6. We conjecture that Theorem 1 holds for B an arbitrary compact in l_2.

Lemma 1. *Suppose that B is a compact subset of l_2 and $y_j = \mu_j + n^{-1/2}\varepsilon_j$, $j \geqslant 1$, where $\boldsymbol{\varepsilon}_p \equiv (\varepsilon_1, \ldots, \varepsilon_p)^{\mathrm{T}}$ is Gaussian with mean 0 covariance Σ_p. Let $\Lambda = \{(\boldsymbol{\mu}_p, \Sigma_p) : \boldsymbol{\mu} \in B_{\mathbf{a},d}, \lambda_{\max}(\Sigma_p) \leqslant k_2 < \infty\}$ for $k_2 \geqslant 1$. Then there exist $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \hat{\mu}_2, \ldots)^{\mathrm{T}}$ such that*

$$\max\{\mathrm{E}_{\boldsymbol{\theta}}\|\hat{\boldsymbol{\mu}}_p - \boldsymbol{\mu}_p\|^2 : \boldsymbol{\theta} = (\boldsymbol{\mu}, \Sigma) \in \Lambda\} = o(1),$$

where $\hat{\boldsymbol{\mu}}_p$ and $\boldsymbol{\mu}_p$ follow the same convention. In fact,

$$\max\{\mathrm{E}_{\boldsymbol{\theta}}\|\hat{\boldsymbol{\mu}}_p - \boldsymbol{\mu}_p\|^2 : \boldsymbol{\theta} \in \Lambda\} \leqslant \mathrm{k}_2 \max\{\mathrm{E}_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 : \boldsymbol{\mu} \in B_{\mathbf{a},d}\}.$$

Proof of Theorem 1. We first prove (a). Suppose $\Sigma = I$. Then $(\hat{\lambda}_1, \ldots, \hat{\lambda}_n)$, $(\hat{\boldsymbol{\xi}}_1, \ldots, \hat{\boldsymbol{\xi}}_n)$ are independent and $\hat{\boldsymbol{\xi}}_j$ are identically distributed uniformly on the unit p-sphere. Moreover, $\hat{\boldsymbol{\Delta}}$ is independent of the $\hat{\lambda}_i$ and $\hat{\boldsymbol{\xi}}_i$. We need to argue that when $\Sigma = I$,

$$\Psi_{\mathrm{I}}(\hat{\boldsymbol{\Delta}}, \hat{\Sigma}^{-}) \xrightarrow{P} 0. \tag{2.8}$$

Write, using the spectral theorem,

$$\Psi_{\mathrm{I}}(\hat{\boldsymbol{\Delta}}, \hat{\Sigma}^{-}) = \frac{\sum_{i=1}^{n} \hat{\lambda}_i^{-1}(\hat{\boldsymbol{\Delta}}, \hat{\boldsymbol{\xi}}_i)^2}{\left(\sum_{i=1}^{n} \hat{\lambda}_i^{-2}(\hat{\boldsymbol{\Delta}}, \hat{\boldsymbol{\xi}}_i)^2\right)^{1/2}}.$$

Use Cauchy–Schwarz and divide the top and bottom by $\sum_{i=1}^{n}(\hat{\boldsymbol{\Delta}}, \hat{\boldsymbol{\xi}}_i)^2$ to obtain

$$\Psi_{\mathrm{I}}(\hat{\boldsymbol{\Delta}}, \hat{\Sigma}^{-}) \leqslant \sqrt{\sum_{i=1}^{n}(\hat{\boldsymbol{\Delta}}, \hat{\boldsymbol{\xi}}_i)^2}.$$

Condition on $\hat{\boldsymbol{\Delta}}$ and take expectations inside the square root to obtain $\sqrt{(n/p)\|\hat{\boldsymbol{\Delta}}\|^2}$. Applying Lemma 1 to $y_j = \bar{X}_{ij}$, $i = 0, 1$, gives $\|\hat{\boldsymbol{\Delta}}\|^2 \xrightarrow{P} \|\boldsymbol{\Delta}\|^2$, and result (a) follows.

We now prove (b). We first argue that under the given condition,

$$\max_{1 \leqslant i \leqslant p} |\hat{\sigma}_{ii}^{-1} - \sigma_{ii}^{-1}| \xrightarrow{P} 0 \tag{2.9}$$

uniformly on Γ where $\hat{\Sigma} \equiv \|\hat{\sigma}_{ij}\|$, $\Sigma = \|\sigma_{ij}\|$. Since $0 < k_1 \leqslant \sigma_{ii} \leqslant k_2 < \infty$ for all Σ such that $\theta \in \Gamma$, (2.9) follows from

$$\max_{1 \leqslant i \leqslant p} |\hat{\sigma}_{ii} - \sigma_{ii}| \xrightarrow{P} 0 \tag{2.10}$$

uniformly on Γ. But by Lemma 4 in Section 6,

$$P\left[\max_{1 \leqslant i \leqslant p}\left|\frac{\hat{\sigma}_{ii}}{\sigma_{ii}} - 1\right| \geqslant \varepsilon\right] \leqslant 2p\mathrm{e}^{-nc(\varepsilon)}$$

for $c(\varepsilon) > 0$. Thus, again invoking $\sigma_{ii} \geqslant k_1 > 0$, (2.10) follows.

Next, let $\delta_{\mathrm{I}}^{\boldsymbol{\Delta}}$ be the rule δ_{I} with $\hat{\boldsymbol{\Delta}}$ replaced by the true $\boldsymbol{\Delta}$. By the monotonicity of $\bar{\Phi}$ on rays,

$$\max\{W(\delta_{\mathrm{I}}^{\boldsymbol{\Delta}}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Gamma\} = \max\{\bar{\Phi}(\Psi_{\Sigma}(\boldsymbol{\Delta}, \hat{D})) : \boldsymbol{\theta} \in \Gamma, \boldsymbol{\Delta}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\Delta} = c^2\}. \tag{2.11}$$

On the other hand,

$$\bar{W}(\delta_{\mathrm{I}}, \boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}\bar{\Phi}(\Psi_{\Sigma}(\hat{\boldsymbol{\Delta}}, \hat{D})),$$

where $\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_0$ given by (2.2). We show in Lemma 5 that

$$\max_{\Gamma} \bar{W}(\delta_{\mathrm{I}}, \boldsymbol{\theta}) = \max\{\bar{W}(\delta_{\mathrm{I}}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Gamma, c^2 \leqslant \boldsymbol{\Delta}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\Delta} \leqslant f^2 < \infty\} \tag{2.12}$$

for some $f^2 < \infty$. Therefore, in view of (2.11), to prove (2.5) we need only check two things:

$$\max\{|\Psi_\Sigma(\hat{\mathbf{\Delta}}, \hat{D}) - \Psi_\Sigma(\mathbf{\Delta}, \tilde{D})| : \boldsymbol{\theta} \in \Gamma, c^2 \leqslant \mathbf{\Delta}^{\mathrm{T}}\Sigma^{-1}\mathbf{\Delta} \leqslant f^2\} = o_p(1), \tag{2.13}$$

where $\tilde{D}$ denotes either D or $\hat{D}$, uniformly on B; and

$$\begin{aligned}&\min\{\Psi_\Sigma(\mathbf{\Delta}, D) : \boldsymbol{\theta} \in \Gamma, c^2 \leqslant \mathbf{\Delta}^{\mathrm{T}}\Sigma^{-1}\mathbf{\Delta} \leqslant f^2\}\\ &\quad = \min\{\Psi_\Sigma(\mathbf{\Delta}, D) : \boldsymbol{\theta} \in \Gamma, \mathbf{\Delta}^{\mathrm{T}}\Sigma^{-1}\mathbf{\Delta} \geqslant c^2\} = \frac{\sqrt{K_0}}{1+K_0}c.\end{aligned} \tag{2.14}$$

To see why this is sufficient to establish result (2.5), note first that by (2.12) we need only consider $\overline{W}_{\mathrm{I}}$ on the set $\tilde{\Gamma}$ where $\boldsymbol{\theta} \in \Gamma$, $c^2 \leqslant \mathbf{\Delta}^{\mathrm{T}}\Sigma^{-1}\mathbf{\Delta} \leqslant f^2$. Next, by (2.13), we can replace δ_{I} in W by $\delta_{\mathrm{I}}^{\mathbf{\Delta}}$ on $\tilde{\Gamma}$. Replacing δ_{I} by $\delta_{\mathrm{I}}^{\mathbf{\Delta}}$ in W implies that this replacement in $\overline{W}$ is also permitted since $0 \leqslant \overline{\Phi} \leqslant 1$. Then (2.11) permits us to consider $W(\delta_{\mathrm{I}}^{\mathbf{\Delta}}, \boldsymbol{\theta})$ and hence $\overline{W}(\delta_{\mathrm{I}}^{\mathbf{\Delta}}, \boldsymbol{\theta})$ just for $\boldsymbol{\theta} \in \Gamma$, $\mathbf{\Delta}^{\mathrm{T}}\Sigma^{-1}\mathbf{\Delta} = c^2$. Again by (2.13), we can replace $\hat{D}$ in $\delta_{\mathrm{I}}^{\mathbf{\Delta}}$ by D.

Now, to verify the second equality in (2.14), note that

$$\begin{aligned}&\min\{\Psi_\Sigma(\mathbf{\Delta}, D) : \mathbf{\Delta}^{\mathrm{T}}\Sigma^{-1}\mathbf{\Delta} = c^2\}\\ &\quad = \frac{c}{2}\min\left\{\frac{\Psi_\Sigma(\mathbf{\Delta}, D)}{\Psi_\Sigma(\mathbf{\Delta}, \Sigma)} : \mathbf{\Delta}^{\mathrm{T}}\Sigma^{-1}\mathbf{\Delta} = c^2\right\}.\end{aligned} \tag{2.15}$$

But the ratio $\Psi_\Sigma(\mathbf{\Delta}, D)/\Psi_\Sigma(\mathbf{\Delta}, \Sigma)$ is invariant under $\Sigma \to b\Sigma$ for any $b > 0$. We conclude that

$$\min\{\Psi_\Sigma(\mathbf{\Delta}, D) : \mathbf{\Delta}^{\mathrm{T}}\Sigma^{-1}\mathbf{\Delta} = c^2\} = \frac{c}{2}r,$$

where r is given by (2.7). Moreover, the bound (2.7) is sharp when all eigenvalues are equal, which establishes (2.14).

To complete the proof of (2.5), we need only check (2.13). In view of (2.9) and Lemma 1, (2.13) will follow from (here $\|\cdot\|$ is the l_2 or operator norm as appropriate)

$$|\Psi_\Sigma(\hat{\mathbf{\Delta}}, \hat{D}) - \Psi_\Sigma(\mathbf{\Delta}, D)| \leqslant C(\|\hat{\mathbf{\Delta}} - \mathbf{\Delta}\| + \|\hat{D} - D\|)$$

for $\|\hat{\mathbf{\Delta}} - \mathbf{\Delta}\| \leqslant \delta_1$, $\|\hat{D} - D\| \leqslant \delta_2$ for δ_1, δ_2 small enough uniformly for $\boldsymbol{\theta} \in \Gamma$, $c^2 \leqslant \mathbf{\Delta}^{\mathrm{T}}\Sigma^{-1}\mathbf{\Delta} \leqslant f^2$ with C depending on $c, f, \mathbf{a}$ only. This is equivalent to the Fréchet derivatives of $\Psi_\Sigma(\mathbf{\Delta}^{\mathrm{T}}, D^{\mathrm{T}})$ being uniformly bounded in a neighbourhood of $(\mathbf{\Delta}, D)$. We shall not argue this here but prove a stronger result (Theorem 2) in Section 4. □

3. Connections to spectral density

If we think of Σ_0 as the covariance of a stationary process $\{\xi_t\}$, the condition on the eigenvalues of the correlation matrix that ensures the good performance of the IR can be related to the spectral density of the corresponding process. In an abuse of the notation, if we write

$$\Sigma_0 = \|\sigma_{ij}\|_{i,j \geqslant 1} = \|\sigma(|i-j|)\|_{i,j \geqslant 1}, \tag{3.1}$$

then we can think of Σ_0 as the correlation matrix for $p = \infty$, with correlations for finite p obtained by taking the first p rows and columns of Σ_0.

In this case, it is known (Grenander and Szegö 1984) that the $\sigma(m)$ have a spectral representation

$$\sigma(m) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{\mathrm{i}m\gamma} \mathrm{d}F(\gamma)$$

for a finite measure F. F is absolutely continuous with density f which is in L_2 if and only if $\Sigma_m \sigma^2(m) < \infty$ and

$$f(\nu) = \frac{1}{2\pi} \sum_{m=-\infty}^{\infty} \mathrm{e}^{\mathrm{i}m\nu} \sigma(m) \tag{3.2}$$

is the spectral density. Moreover,

$$\lambda_{\max}(\Sigma_0) = \sup_{\nu} f(\nu),$$

$$\lambda_{\min}(\Sigma_0) = \inf_{\nu} f(\nu).$$

In particular, any process with the spectral density bounded by positive constants

$$0 < M^{-1} \leqslant f(\nu) \leqslant M < \infty, \qquad \text{for all } \nu, \tag{3.3}$$

would have a covariance function that satisfies our constraints.

Note that $\xi_t = X_t/\sigma_t$ is the stationary process here, and not the original set of variables X_t, which are still allowed to have different variances. The assumption of stationarity is not necessarily realistic for classification problems, but the connection to spectral density provides a useful tool for investigating some examples below.

Example 1. Let $\{X_t\}$ be an ARMA(r, q) process defined by

$$\phi(B)X_t = \theta(B)Z_t$$

where B is the shift operator,

$$\phi(z) = 1 - \phi_1 z - \ldots - \phi_r z^r, \qquad \theta(z) = 1 + \theta_1 z + \ldots \theta_q z^q,$$

and $\{Z_t\}$ is a white noise process with variance σ^2. Then as long as $\phi(z)$ has no zeros on the unit circle and no common zeros with $\theta(z)$, X_t has spectral density

$$f(\nu) = \frac{\sigma^2}{2\pi} \frac{|\theta(\mathrm{e}^{\mathrm{i}\nu})|^2}{|\phi(\mathrm{e}^{-\mathrm{i}\nu})|^2},$$

which satisfies the constraint (3.3) whenever both $\phi(z)$ and $\theta(z)$ have no zeros on the unit circle.

In particular, the AR(1) process, which corresponds to

$$\Sigma_0 = \|\sigma(|i-j|)\|_{i,j \geqslant 1} = \|\rho^{|i-j|}\|_{i,j \geqslant 1},$$

has $\theta(z) \equiv 1$, $\phi(z) = 1 - \rho z$, $|\rho| < 1$, so for this form of covariance matrix the IR rule result holds. In this case one can also compute

$$K = \frac{\sup_{\nu} f(\nu)}{\inf_{\nu} f(\nu)} = \frac{(1+\rho)^2}{(1-\rho)^2}.$$

Similarly, the MA(1) process, which corresponds to a tridiagonal correlation matrix with $\sigma(1) = \rho$, $|\rho| < 0.5$, has $\phi(z) \equiv 1$, $\theta(z) = 1 + \rho z$, so this type of covariance structure also benefits from using the IR. Here also $K = (1+\rho)^2/(1-\rho)^2$. These examples should be viewed primarily as motivational, though such covariance structures may occur in classification of time series data or data generated by a stationary random field, which is a reasonable model for some types of image data – in particular, for texture.

Example 2. A simple example where condition (3.3) is not satisfied is provided by the correlation matrix

$$\Sigma_0 = \|\sigma_{ij}\| = \begin{cases} 1, & \text{if } i = j, \\ \rho, & \text{if } i \neq j. \end{cases}$$

This corresponds to the process $X_t = X_0 + \varepsilon_t$, $\{\varepsilon_t\}$ white noise, for which the spectral density does not exist. One can also check directly that the eigenvalues of its $p \times p$ subsection are $\lambda_1 = \ldots = \lambda_{p-1} = 1 - \rho$, $\lambda_p = 1 + (p-1)\rho$, so that $\lambda_{\max}(\Sigma_0)/\lambda_{\min}(\Sigma_0) \to \infty$ as $p \to \infty$ and the worst-case error of the IR is also $\frac{1}{2}$.

Necessary and sufficient conditions for the spectral density to be bounded between two positive constants were given by Bradley (2002), in terms of what he called 'linear dependence coefficients' of the process. While these conditions are not in general easy to check, they may be useful in special cases.

4. The Gaussian 'coloured' noise model and Bayes consistency and minimax regret

To motivate and justify rules which interpolate between δ_{F} and δ_{I}, we need an asymptotic framework which permits us to make $\Sigma_{p\times p}$ converge as $n \to \infty$. We make our discussion more systematic. The Gaussian coloured noise model is given by (see, for example, Johnstone 2002)

$$\mathbf{X}_i = \boldsymbol{\mu}_i + n^{-1/2}\boldsymbol{\varepsilon}, \qquad i = 0, 1,$$

where $\boldsymbol{\mu}_i \in l_2$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots)$ is a sequence of Gaussian variables with mean 0 and $\mathrm{cov}(\varepsilon_a, \varepsilon_b) = \sigma_{ab}$, $1 \leqslant a \leqslant b < \infty$. Let Σ_p be the upper $p \times p$ corner of Σ, that is, $\|\mathrm{cov}(\varepsilon_i, \varepsilon_j)\|$, $1 \leqslant i, j \leqslant p$. We denote $(\mathbf{a}, \mathbf{b}) \equiv \sum_{i=1}^{\infty} a_i b_i$ for $\mathbf{a} = (a_1, a_2, \ldots)$ and $\mathbf{b} = (b_1, b_2, \ldots)$ and $\|\mathbf{a}\| = \sum_{i=1}^{\infty} a_i^2$ as usual. Now $\Sigma \equiv \|\sigma_{ij}\|$, $1 \leqslant i, j < \infty$, is an infinite-dimensional matrix. Suppose Σ can be viewed as a linear operator from l_2 to l_2: if $\mathbf{a} \in l_2$,

$$\Sigma\mathbf{a} = \mathbf{b} \in l_2 \tag{4.1}$$

where $b_i = \sum_{j=1}^{\infty} \sigma_{ij} a_j$. This holds if and only if $\sum_{j=1}^{\infty} \sigma_{ij}^2 < \infty$ for all i. We assume that Σ is bounded and has a bounded inverse, that is, for all $\mathbf{a} \in l_2$,

(i) $(\mathbf{a}, \Sigma\mathbf{a}) \leqslant M\|\mathbf{a}\|^2$,
(ii) $(\mathbf{a}, \Sigma\mathbf{a}) \geqslant M^{-1}\|\mathbf{a}\|^2$,

for some M finite. Such a Σ is a Toeplitz operator, since it satisfies $\sigma_{ij} = \sigma(|i-j|)$. If $O = \|o_{ij}\|_{1 \leqslant i,j < \infty}$ is a linear operator from l_2 to l_2 operating as in (4.1), then its operator norm is given by

$$\|O\| \equiv \sup\{\|O\mathbf{a}\| : \|\mathbf{a}\| = 1\}.$$

If O is symmetric (self-adjoint), $o_{ij} = o_{ji}$ for all i, j, then it is well known that the spectrum of O is real and discrete, $\lambda_1(O)$, $\lambda_2(O)$, ... and

$$\|O\| = \sup_j |\lambda_j(O)| = \max\{|\lambda_{\max}(O)|, |\lambda_{\min}(O)|\}. \tag{4.2}$$

It follows that, for Σ as above,

$$\|\Sigma\| = \lambda_{\max}(\Sigma) \leqslant M,$$

$$\|\Sigma^{-1}\| = \lambda_{\max}(\Sigma^{-1}) = \lambda_{\min}^{-1}(\Sigma) \leqslant M. \tag{4.3}$$

For a Toeplitz operator, one can show more than (4.3). We summarize the facts we need below as Lemma 2, and refer to Grenander and Szegö (1984) for proof; see also Böttcher *et al.* (1996).

Lemma 2. *Suppose T is a linear operator from l_2 to l_2 which is self-adjoint and Toeplitz,*

$$t_{ij} = t(i-j), \quad t(j) = t(-j), \qquad \text{all } j.$$

If $\sum_{k=0}^{\infty} t^2(k) < \infty$, then

$$g_T(x) = \sum_{k=-\infty}^{\infty} \mathrm{e}^{\mathrm{i}kx} t(k)$$

is in $L_2(-\pi, \pi)$ and

$$\|T\| = \sup_x |g_T(x)|; \tag{4.4}$$

and if T^{-1} is bounded, then

$$\|T^{-1}\| = \left(\inf_x |g_T(x)|\right)^{-1}. \tag{4.5}$$

Thus, if $\sum_k t^2(k) < \infty$, conditions (i) and (ii) are equivalent to (3.3).

The class of Toeplitz operators corresponding to spectral densities satisfying (3.3) suggest rules interpolating between δ_{F} and δ_{I}. We define $\delta_{\mathrm{I}d}$ as the rule which replaces $\hat{\Sigma}_p$ by $\hat{\Sigma}_p^{(d)}$ given below:

$$\mathrm{I}_d\mathrm{R} : \hat{\Sigma}_p^{(d)} = \|\tilde{\sigma}_{ab}^{(d)}\|_{p\times p}, \tag{4.6}$$

$$\tilde{\sigma}_{ab}^{(d)} = \begin{cases} \tilde{\sigma}(b-a), & |a-b| \leqslant d \\ 0 & \text{otherwise,} \end{cases}$$

$$\tilde{\sigma}(k) = (p-k)^{-1}\left\{\sum_{a=1}^{p-k} \hat{\sigma}_{a,a+k}\right\}.$$

The rules $\mathrm{I}_d\mathrm{R}$ are natural if we assume that Σ_p is the covariance matrix of p consecutive observations from a moving average of order $d+1$. Let $\boldsymbol{\theta} = (\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \Sigma)$, where $\boldsymbol{\mu}_i$ range over subsets of l_2 and Σ over a subset of Toeplitz operators with spectral densities satisfying (3.3) and smoothness restrictions. Let

$$R_\Gamma(\delta) = \overline{W}_\Gamma(\delta) - \overline{\Phi}\left(\frac{c}{2}\right),$$

the difference between the maximum and minimax risks, sometimes called the regret of δ. Let $R_\Gamma \equiv R_\Gamma(\delta_{\mathrm{I}d_n})$, where we suppress dependence on p and n in R_Γ. Define

$$\Gamma_r = \{\boldsymbol{\theta} : \boldsymbol{\mu}_\mathrm{i} \in \mathrm{B}_{\mathbf{a},d},\ M^{-1} \leqslant f_\Sigma \leqslant M < \infty,\ (\Sigma^{-1}\boldsymbol{\Delta}, \boldsymbol{\Delta}) \geqslant c^2,\ \|f_\Sigma^{(r)}\|_\infty \leqslant M_r\},$$

where $f_\Sigma^{(r)}$ is the rth derivative of f_Σ. Suppose that $n^{-\alpha}$ is the rate for estimating $\boldsymbol{\mu}$ when Σ is the identity (Gaussian white noise), that is,

$$\max\{\mathrm{E}_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 : \boldsymbol{\mu} \in B_{\mathbf{a},d}\} \asymp n^{-\alpha},\ \alpha < 1, \tag{4.7}$$

and let

$$\gamma = \min\left\{\alpha, \frac{2r}{2r+1}\right\}.$$

Theorem 2. *There exist $d_n \to \infty$ (dependent only on $\mathbf{a}$, r) such that*

$$R_{\Gamma_r} \asymp n^{-\gamma}\Omega(n), \tag{4.8}$$

where $\Omega(n) = O(\log n)$.

We give d_n below, and conjecture that $n^{-\gamma}\Omega(n)$ in fact has the minimax property that

$$\min_\delta \overline{W}_\Gamma(\delta) - \overline{\Phi}\left(\frac{c}{2}\right) \asymp n^{-\gamma}\Omega(n). \tag{4.9}$$

Proof of Theorem 2. We will write $\boldsymbol{\Delta}_p$ to signify the first p coordinates of $\boldsymbol{\Delta} \in l_2$. We claim that, for all $\theta \in \Gamma_r$ with $c^2 \leqslant \boldsymbol{\Delta}_p^\mathrm{T}\Sigma_p^{-1}\boldsymbol{\Delta}_p \leqslant f^2$,

$$|\Psi_{\Sigma_p}(\boldsymbol{\Delta}_p^*, \Sigma_p^*) - \Psi_{\Sigma_p}(\boldsymbol{\Delta}_p, \Sigma_p)| \leqslant C\{\|\boldsymbol{\Delta}_p^* - \boldsymbol{\Delta}_p\|^2 + \|\Sigma_p^* - \Sigma_p\|^2\}, \tag{4.10}$$

for all $\|\boldsymbol{\Delta}_p^* - \boldsymbol{\Delta}_p\| \leqslant \delta_1$, $\|\Sigma_p^* - \Sigma_p\| \leqslant \delta_2$, for δ_1, δ_2 small enough and C depending on Γ_r and f^2 only. The bounds are valid for $p = \infty$ as well.

This is equivalent to showing that, if D, D^2 denote Fréchet derivatives,

(a) $D\Psi_{\Sigma_p}(\mathbf{\Delta}_p, \Sigma_p) = 0$,
(b) $\sup\{|D^2\Psi_{\Sigma_p}(\mathbf{\Delta}_p^*, \Sigma_p^*)|] : \|\mathbf{\Delta}_p^* - \mathbf{\Delta}_p\| \leqslant \delta_1, \|\Sigma_p^* - \Sigma_p\| \leqslant \delta_2\} \leqslant C < \infty$ for all $\boldsymbol{\theta}$ as above.

Until we need them again, we shall drop the p subscripts.

To show (a), we expand the numerator of Ψ as

$$(\mathbf{\Delta} + \lambda_1\mathbf{e})^{\mathrm{T}}(\Sigma^{-1} + \lambda_2 E)(\mathbf{\Delta} + \lambda_1\mathbf{e}) = \mathbf{\Delta}^{\mathrm{T}}\Sigma^{-1}\mathbf{\Delta} + 2\lambda_1\mathbf{e}^{\mathrm{T}}\boldsymbol{\Sigma}^{\mathrm{T}}\mathbf{\Delta} + \lambda_2\mathbf{\Delta}^{\mathrm{T}}E\mathbf{\Delta} + O(\lambda_1 + \lambda_2)^2, \tag{4.11}$$

and the denominator of Ψ as

$$\begin{aligned}&(\mathbf{\Delta} + \lambda_1\mathbf{e})^{\mathrm{T}}(\Sigma^{-1} + \lambda_2 E)\Sigma(\Sigma^{-1} + \lambda_2 E)(\mathbf{\Delta} + \lambda_1\mathbf{e})\\ &\qquad = (\mathbf{\Delta}^{\mathrm{T}}\Sigma^{-1}\mathbf{\Delta} + 2\lambda_2\mathbf{\Delta}^{\mathrm{T}}E\mathbf{\Delta} + 2\lambda_1\mathbf{e}^{\mathrm{T}}\Sigma^{-1}\mathbf{\Delta}) + O(\lambda_1 + \lambda_2)^2)^{1/2}\\ &\qquad = (\mathbf{\Delta}^{\mathrm{T}}\Sigma^{-1}\mathbf{\Delta})^{1/2}(1 + \lambda_1\mathbf{e}^{\mathrm{T}}\Sigma^{-1}\mathbf{\Delta} + \lambda_2\mathbf{\Delta}^{\mathrm{T}}E\mathbf{\Delta}) + O(\lambda_1 + \lambda_2)^2.\end{aligned} \tag{4.12}$$

Hence,

$$\Psi_\Sigma(\mathbf{\Delta} + \lambda_1\mathbf{e}, (\Sigma^{-1} + \lambda_2 E)^{-1}) - \Psi_\Sigma(\mathbf{\Delta}, \Sigma) = (\mathbf{\Delta}^{\mathrm{T}}\Sigma^{-1}\mathbf{\Delta})^{1/2}O(\lambda_1 + \lambda_2)^2$$

and (a) follows.

For (b), it is clear that we need to bound terms appearing in (4.11) from above and in (4.12) from below uniformly on the specified set for $|\lambda_1| \leqslant \delta_1$, $|\lambda_2| \leqslant \delta_2$, $\|\mathbf{e}\| \leqslant 1$, $\|E\| \leqslant 1$. The upper bounds are straightforward. For instance,

$$\begin{aligned}[\mathbf{\Delta}^*]^{\mathrm{T}}E\mathbf{\Delta}^* &\leqslant \|\mathbf{\Delta}^*\|^2 \leqslant 2(\|\mathbf{\Delta}\|^2 + \delta_1^2)\\ &\leqslant 2(k_2 f^2 + \delta_1^2),\end{aligned}$$

since $\|\mathbf{\Delta}\|^2/k_2 \leqslant \mathbf{\Delta}^{\mathrm{T}}\Sigma^{-1}\mathbf{\Delta} \leqslant f^2$.

On the other hand,

$$\begin{aligned}\lambda_{\min}([\Sigma^*]^{-1}\Sigma[\Sigma^*]^{-1}) = \frac{1}{\lambda_{\max}(\Sigma^*\Sigma^{-1}\Sigma^*)} &\geqslant \frac{1}{\|\Sigma^*\|^2\|\Sigma^{-1}\|}\\ &\geqslant \frac{k_1}{(\|\Sigma\| + \delta_2)^2} \geqslant \frac{k_1}{(k_2 + \delta_2)^2}.\end{aligned}$$

Hence,

$$\begin{aligned}[\mathbf{\Delta}^*]^{\mathrm{T}}[\Sigma^*]^{-1}\Sigma[\Sigma^*]^{-1}[\mathbf{\Delta}^*] &\geqslant \frac{k_1}{(k_2 + \delta_2)^2}\|\mathbf{\Delta}^*\|^2\\ &\geqslant \frac{k_1}{(k_2 + \delta_2)^2}(\|\mathbf{\Delta}\| - \delta_1)^2 \geqslant \frac{k_1}{(k_2 + \delta_2)^2}(c\sqrt{k_1} - \delta_1)^2,\end{aligned} \tag{4.13}$$

which is bounded away from 0 for δ_1 small. Claim (b) follows.

From (4.10), we see that

$$
\begin{aligned}
&|\mathrm{E}_{\boldsymbol{\theta}}\overline{\Phi}(\Psi_{\Sigma_p}(\hat{\boldsymbol{\Delta}}_p, \hat{\Sigma}_p^{(d)})) - \mathrm{E}_{\boldsymbol{\theta}}\overline{\Phi}(\Psi_{\Sigma_p}(\boldsymbol{\Delta}_p, \Sigma_p^{(d)}))| \\
&\quad \leqslant C(\mathrm{E}_{\boldsymbol{\theta}}\|\hat{\boldsymbol{\Delta}}_p - \boldsymbol{\Delta}_p\|^2 1(\|\hat{\boldsymbol{\Delta}}_p - \boldsymbol{\Delta}_p\| \leqslant \delta_1) + \mathrm{E}_{\boldsymbol{\theta}}\|\hat{\Sigma}_p^{(d)} - \Sigma_p\|^2 1(\|\hat{\Sigma}_p - \Sigma_p\| \leqslant \delta_2) \\
&\qquad + P[\|\hat{\boldsymbol{\Delta}}_p - \boldsymbol{\Delta}_p\| \geqslant \delta_1] + P[\|\hat{\Sigma}_p^{(d)} - \Sigma_p\| \geqslant \delta_2]). \qquad (4.14)
\end{aligned}
$$

Let X_{kj} denote the jth component of observation $\mathbf{X}_k$, $k = 1, \ldots, n$. Write, taking $\boldsymbol{\mu} = 0$,

$$
\tilde{\sigma}(a) = \frac{1}{n(p-a)}\sum_{k=1}^{n}\sum_{j=1}^{p-a} X_{kj}X_{k(j+a)} = \frac{1}{(p-a)}\sum_{j=1}^{p-a}\overline{X}_j\overline{X}_{(j+a)},
$$

where $\overline{X}_j = (1/n)\sum_{k=1}^{n} X_{kj}$. We use Lemma 4 to bound, for $a \geqslant 1$, $P[|\tilde{\sigma}(a) - \sigma(a)| \geqslant \nu]$. We can apply Lemma 4 since $(X_{k1}, \ldots X_{kp})$ are independent and identically distributed as $N(\mathbf{0}, \Sigma_p)$, and $\sqrt{n}(\overline{X}_1, \ldots, \overline{X}_p)$ are $N(\mathbf{0}, \Sigma_p)$ as well.

By Lemma 4,

$$
\begin{aligned}
P[\max\{|\hat{\sigma}(a) - \sigma(a)| : |a| \leqslant d\} \geqslant \nu] &\leqslant d\max\{P[|\hat{\sigma}(a) - \sigma(a)| \geqslant \nu] : |a| \leqslant d\} \\
&\leqslant dK_1 \exp\{-n(p-d)c_1(\nu)\}
\end{aligned}
$$

Suppose $d \leqslant p/2$. Then, for some $\varepsilon > 0$, $A < \infty$, $\delta > 0$, $\delta \geqslant \nu \geqslant A/\sqrt{np}$,

$$
dK_1 \exp\{-n(p-d)c_1(\nu)\} \leqslant \mathrm{e}^{-\varepsilon np\nu^2} \qquad (4.15)
$$

since $c_1(\nu) \geqslant b_1\nu^2$, for $\nu \leqslant \delta$ sufficiently small.

Let

$$
V_d \equiv \max\{|\tilde{\sigma}(a) - \sigma(a)| : |a| \leqslant d\}.
$$

By Lemma 6,

$$
V_d \leqslant \|\hat{\Sigma}_p^{(d)} - \Sigma_p^{(d)}\| \leqslant (2\sqrt{pd} + 1)V_d. \qquad (4.16)
$$

Therefore, by (4.16),

$$
\begin{aligned}
\mathrm{E}_{\boldsymbol{\theta}}\|\hat{\Sigma}_p^{(d)} - \Sigma_p^{(d)}\|^2 1(\|\hat{\Sigma}_p^{(d)} - \Sigma_p^{(d)}\| \leqslant \delta) &\leqslant 4pd\int_0^{\delta} \nu P[V_d \geqslant \nu]\,\mathrm{d}\nu \\
&\leqslant 4pd\left(\frac{A^2}{2np} + \int_{\sqrt{A/np}}^{\infty} \nu\mathrm{e}^{-\varepsilon np\nu^2}\,\mathrm{d}\nu\right) \\
&\leqslant B_1\frac{d}{n}, \qquad (4.17)
\end{aligned}
$$

for some universal B_1. On the other hand,

$$
P[\|\hat{\Sigma}_p^{(d)} - \Sigma_p^{(d)}\| \geqslant \delta] \leqslant B_2\mathrm{e}^{-n\varepsilon} \leqslant B_2\frac{d}{n}, \qquad (4.18)
$$

for $\log n < \varepsilon n$. Concluding, we see from (4.14)–(4.18) and assumption (4.7) that if $d \leqslant p/2$, then

$$\sup\{|\mathrm{E}_{\boldsymbol{\theta}}(\overline{\Phi}(\Psi_{\Sigma_p}(\hat{\boldsymbol{\Delta}}_p, \hat{\Sigma}_p^{(d)})) - \overline{\Phi}(\Psi_{\Sigma_p}(\boldsymbol{\Delta}_p, \Sigma_p^{(d)})| : c^2 \leqslant \boldsymbol{\Delta}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\Delta} \leqslant f^2, \boldsymbol{\theta} \in \Gamma_r\}$$

$$\leqslant C \max\left\{\frac{d}{n}, n^{-\alpha}\right\}. \tag{4.19}$$

Now we appeal to Lemma 7, which yields that

$$\|\Sigma_p^{(d)} - \Sigma_p\| \leqslant \frac{B \log d}{d^r}, \tag{4.20}$$

$$\|\Sigma_p - \Sigma\| \leqslant \frac{B \log p}{p^r}. \tag{4.21}$$

Therefore, if $d \to \infty$, $\hat{\Sigma}_p^{(d)}$ satisfy the conditions of Lemma 5, and we can conclude that

$$\overline{W}_{\Gamma_r}(\delta_{\mathrm{I}d}) = \sup\{\mathrm{E}_{\boldsymbol{\theta}}\overline{\Phi}\Psi_{\Sigma_p}(\hat{\boldsymbol{\Delta}}_p, \hat{\Sigma}_p^{(d)}) : \boldsymbol{\theta} \in \Gamma_r, c^2 \leqslant \boldsymbol{\Delta}_p^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\Delta}_p \leqslant f^2\}$$

$$+ O(P[\|\hat{\Sigma}_p^{(d)} - \Sigma_p^{(d)}\| \geqslant \delta]). \tag{4.22}$$

Putting (4.19)–(4.22) together, we obtain that if $d \leqslant p/2$,

$$\overline{W}_{\Gamma_r}(\delta_{\mathrm{I}d}) = \sup\left\{\overline{\Phi}\left(\frac{1}{2}(\boldsymbol{\Delta}_p^{\mathrm{T}}\Sigma_p^{-1}\boldsymbol{\Delta}_p)^{1/2}\right) : \boldsymbol{\Delta}_p^{\mathrm{T}}\Sigma_p^{-1}\boldsymbol{\Delta}_p \geqslant c^2\right\}$$

$$+ O\left(\max\left\{\frac{d}{n}, n^{-\alpha}, \frac{(\log d)^{2r}}{d^{2r}}\right\}\right)$$

$$= \overline{\Phi}\left(\frac{c}{2}\right) + O(n^{-\gamma}\log n) \tag{4.23}$$

by taking p sufficiently large, $d = [n(\log n)^{2r}]^{1/(2r+1)}$. The theorem follows. □

5. Discussion

Donoho *et al.* (1995) have remarked that the phenomenon of minimax performance in the presence of large p can occur. By assuming 'sparsity', only a few parameters need to be estimated. Most are nearly zero and should be estimated as zero. A similar phenomenon appears to be occurring here, since the estimates $\hat{\Sigma}_p^{(d)}$ make most of the covariances 0. However, the structure is rather different and clearly the stationary structure plays a major role. We conjecture that other regularity features in the covariance structure more appropriate in higher-dimensional settings, such as the texture case (Levina 2002), can also be taken advantage of. Nevertheless, it is clear that such features can also be viewed as 'sparsity' in an appropriate representation. For instance, in our case, this corresponds to the Fourier series representation of the spectral density and the implicit assumption that higher-order Gaussian coefficients can be neglected. Greenshtein and Ritov (2004) propose to take advantage of the sparsity of $\Sigma^{-1}\boldsymbol{\Delta}$ in another way. Whether their methods will yield minimax results in our context is unclear.

6. Proofs of necessary lemmas

***Proof of Lemma 1 for* $\boldsymbol{\Sigma}$ *general*.** Note that in the case $\Sigma = I$ (Johnstone 2002),

$$\hat{\mu}_i = y_i(1 - r_{in})_+ \tag{6.1}$$

where $x_+ = \max(x, 0)$ and

$$\frac{1}{n}\sum_{i=1}^{\infty}(1 - r_{in})_+^2 \to 0,$$

$$\max\left\{\sum_{i=1}^{\infty}(1 - (1 - r_{in})_+)^2 \mu_i^2 : \boldsymbol{\mu} \in B\right\} \to 0.$$

For arbitrary Σ_p, let $[\tau_{ij}] = \Sigma_p^{1/2}$ and estimate μ_i by (6.1). Then,

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 = \sum_{i=1}^{\infty}(1 - r_{in})_+^2 \left(\sum_{j=1}^{\infty}\tau_{ij}\delta_j\right)^2 + \sum_{i=1}^{\infty}(1 - (1 - r_{in})_+)^2 \mu_i^2$$

where the δ_j are independent and identically distributed as $N(0, 1/n)$. Thus,

$$\mathrm{E}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 = \frac{1}{n}\sum_{i=1}^{\infty}(1 - r_{in})_+^2 \sum_{j=1}^{\infty}\tau_{ij}^2 + \sum_{i=1}^{\infty}(1 - (1 - r_{in})_+)^2 \mu_i^2.$$

Note that

$$\begin{aligned}
\max_i \sum_{j=1}^{p}\tau_{ij}^2 &= \max_i \max\left\{\left(\sum_{j=1}^{p}\tau_{ij}\mu_j\right)^2 : \|\boldsymbol{\mu}\|^2 = 1\right\} \\
&= \max\left\{\max_i\left(\sum_{j=1}^{p}\tau_{ij}\mu_j\right)^2 : \|\boldsymbol{\mu}\| = 1\right\} \\
&\leqslant \max\left\{\sum_{i=1}^{p}\left(\sum_{j=1}^{p}\tau_{ij}\mu_j\right)^2 : \|\boldsymbol{\mu}\| = 1\right\} \\
&\leqslant k_2,
\end{aligned}$$

since $\|\Sigma_p^{1/2}\| = \|\Sigma_p\|^{1/2} \leqslant k_2^{1/2}$. The lemma follows. □

Lemma 3. *Suppose* $0 < \lambda_1 \leqslant \gamma_1 \leqslant \ldots \leqslant \gamma_p \leqslant \lambda_2 < \infty$ *and* $\sum_{j=1}^{p}\gamma_j = p$. *Then, for a suitable* $\epsilon > 0$,

$$M(\nu) = \max\left\{\sum_{j=1}^{p}[\log(1-2\gamma_j s) + 2s\gamma_j(\nu+1)] : 0 \leqslant s \leqslant \frac{1}{2\lambda_2}\right\}$$

$$\geqslant \min\left\{\frac{p\lambda_1^2\epsilon^2\nu^2}{2\lambda_2^3},\ \frac{p(1-\epsilon)\nu}{2\lambda_2}\right\}.$$

Proof. Note that if $0 \leqslant x \leqslant 1-\epsilon$,

$$\log(1-x) + x(\nu+1) \geqslant -\frac{x^2}{2\epsilon^2} + \nu x.$$

Therefore,

$$M(\nu) \geqslant \max\left\{-\frac{\upsilon^2}{2\epsilon^2}\sum_{j=1}^{p}\gamma_j^2 + \upsilon\nu\sum_{j=1}^{p}\gamma_j : \upsilon \leqslant \frac{1-\epsilon}{\lambda_2}\right\}.$$

Substituting

$$\upsilon = \nu\frac{\lambda_1}{\lambda_2}\frac{\sum_{j=1}^{p}\gamma_j}{\sum_{j=1}^{p}\gamma_j^2}\epsilon^2 \leqslant \frac{\nu\epsilon^2}{\lambda_2} \leqslant \frac{1-\epsilon}{\lambda_2}$$

for $\nu \leqslant (1-\epsilon)/\epsilon^2$, we obtain

$$M(\nu) \geqslant \nu^2\frac{\epsilon^2\lambda_1^2}{2\lambda_2^2}\frac{\left(\sum_{j=1}^{p}\gamma_j\right)^2}{\sum_{j=1}^{p}\gamma_j^2} \geqslant \nu^2\frac{\epsilon^2}{2}\frac{\lambda_1^2}{\lambda_2^3}p.$$

On the other hand, for any $\epsilon > 0$,

$$M(\nu) \geqslant p\left[\log\epsilon + \frac{1}{\lambda_2}(1-\epsilon)(\nu+1)\right] \geqslant \frac{p(1-\epsilon)\nu}{2\lambda_2}$$

for $\nu \geqslant 2\lambda_2(\log\epsilon^{-1})(1-\epsilon)^{-1}$ by taking $s = (1-\epsilon)/(2\lambda_2)$. The lemma follows by taking ϵ so that $2\lambda_2(\log\epsilon^{-1})(1-\epsilon)^{-2} \leqslant \epsilon^{-2}$. □

Lemma 4. *Let* $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ *be independent, identically distributed* p*-variate Gaussian with mean* $\mathbf{0}$ *such that* $\mathrm{var}(\mathbf{Z}_1) = \Sigma = \|\sigma(a, b)\|_{p\times p}$, $\sigma(a, a) = 1$ *for all* a, *and* $0 < \lambda_1 \leqslant \lambda_{\min}(\Sigma) \leqslant \lambda_{\max}(\Sigma) \leqslant \lambda_2 < \infty$. *Then*

$$P\left[\left|\sum_{i=1}^{n}\sum_{j=1}^{p}(Z_{ij}^2 - 1)\right| > np\nu\right] \leqslant \exp\{-npc_0(\nu, \lambda_1, \lambda_2)\}.$$

If, further, $\sigma(a, b) = \sigma(|b-a|)$*, then, for all* t*,*

$$P\left[\left|\sum_{i=1}^{n}\sum_{j=1}^{p-t}(Z_{ij}Z_{i(j+t)} - \sigma(t))\right| > n(p-t)\nu\right] \leqslant K_1 \exp\{-n(p-t)c_1(\nu, \lambda_1, \lambda_2)\}.$$

Here, for $m = 0, 1,$

$$c_m(\nu, \lambda_1, \lambda_2) \equiv \min\{a_m(\lambda_1, \lambda_2)\nu,\ b_m(\lambda_1, \lambda_2)\nu^2\},$$

and a_m*,* b_m *are positive functions.*

Proof. We consider the case $t = 0$ and general Σ first. By the spectral theorem,

$$\sum_{i=1}^{n}\sum_{j=1}^{p} Z_{ij}^2 = \sum_{i=1}^{n}\sum_{j=1}^{p}\gamma_j U_{ij}^2,$$

where $\lambda_1 \leqslant \gamma_1 \leqslant \ldots \leqslant \gamma_p \leqslant \lambda_2$ are the eigenvalues of Σ and the U_{ij} are independent $N(0, 1)$.

By Markov's inequality,

$$P\left[\sum_{i=1}^{n}\sum_{j=1}^{p}\gamma_j U_{ij}^2\right] > np(\nu+1) \leqslant \min\left\{\left[\prod_{j=1}^{p}(1-2\gamma_j s)^{-1/2}\mathrm{e}^{-s(\nu+1)}\right]^n : |s| < \frac{1}{2\lambda_2}\right\}.$$

Since $\sum_{j=1}^{p}\gamma_j = p$ by hypothesis, we can apply the bound of Lemma 3. Apply a similar argument to

$$P\left[-\sum_{i=1}^{n}\sum_{j=1}^{p}(Z_{ij}^2 - 1) > np\nu\right]$$

and the first bound follows.

Now write

$$\sum_{i=1}^{n}\sum_{j=1}^{p-t}(Z_{ij}Z_{i(j+t)} - \sigma(t))$$

$$= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{p-t}\{(Z_{ij} + Z_{i(j+t)})^2 - 2(1+\sigma(t)) - (Z_{ij}^2 - 1) - (Z_{i(j+t)}^2 - 1)\}.$$

Further, write the first term as the sum of two terms

$$\sum_{i=1}^{n}\sum\{(Z_{ij} + Z_{i(j+t)})^2 - (1+\sigma(t)) : j \in I_1\} + \sum_{i=1}^{n}\sum\{(Z_{ij} + Z_{i(j+t)})^2 - (1+\sigma(t)) : j \in I_2\},$$

where $j \in I_1 \Leftrightarrow j + t \in I_2$ and the cardinalities of I_1 and I_2 are approximately $(p-t)/2$, that is, differ by at most 1.

Consider the Gaussian vectors $\mathbf{W}_i^{(1)}, \mathbf{W}_i^{(2)}$ corresponding to $(Z_{ij} + Z_{i(j+t)})(2(1+\sigma(t)))^{-1/2}$, $j \in I_1, I_2$ respectively. Call either of these $\mathbf{W}$. Then $\mathbf{W}$ has mean $\mathbf{0}$

and variance covariance matrix of the form $\Gamma = \|\gamma(i, j)\|_{l\times l}$ where $l = (p-t)/2$ or $(p-t+1)/2$ and $(p-t-1)/2$ and $\gamma(0) \equiv 1$.

The maximal eigenvalue of Γ is given by

$$\begin{aligned}\lambda_{\max}(\Gamma) &= \max\{\operatorname{var}(\mathbf{a}^{\mathrm{T}}\mathbf{W}) : \|\mathbf{a}\|^2 = 1\} = \max\left\{\operatorname{var}\sum_{j=1}^{l} b_j W_j) : \sum_{j=1}^{l} b_j^2 = 1\right\}\\ &\leqslant \max\left\{[2(1+\sigma(t)]^{-1}\operatorname{var}\left(\sum_{k=1}^{l} d_k(Z_{1k} + Z_{1(k+t)})\right) : k \in I_1 \text{ or } I_2, \sum_{k=1}^{l} d_k^2 = 1\right\}\\ &\leqslant [2(1+\sigma(t))]^{-1}\max\left\{\operatorname{var}\left(\sum_{j=1}^{p} b_j Z_{1j}\right) : \sum_{j=1}^{p} b_j^2 = 2\right\}\\ &= (1+\sigma(t))^{-1}\lambda_{\max}(\Sigma) \leqslant (1+\sigma(t))^{-1}\lambda_2 \leqslant \frac{\lambda_2}{\lambda_1},\end{aligned}$$

since $(1+\sigma(t)) = \operatorname{var}((Z_{11} + Z_{1t})/\sqrt{2}) \geqslant \lambda_1$.

We obtain a new bound,

$$\begin{aligned}&P\left[\sum_{i=1}^{n}\sum_{j=1}^{p-t}(Z_{ij}Z_{i(j+t)} - \sigma(t)) > n(p-t)\nu\right]\\ &\quad\leqslant P\left[\sum_{i=1}^{n}\sum_{j=1}^{p-t}(Z_{ij}^2 - 1) > \frac{1}{4}n(p-t)\nu\right]\\ &\quad\quad + P\left[\sum_{i=1}^{n}\sum_{j=1}^{p-t}(Z_{i(j+t)}^2 - 1) > \frac{1}{4}n(p-t)\nu\right]\\ &\quad\quad + 2P\left[\sum_{i=1}^{n}\sum_{j=1}^{l}(W_{ij}^2 - 1) > \frac{1}{8}n(p-t)\nu(1+\sigma(t))^{-1}\right],\end{aligned}$$

where we treat W_{ij} generally as components of independent vectors $\mathbf{W}_i$. Now, use $(1+\sigma(t))^{-1} \geqslant 1/\lambda_2$. Apply the $t = 0$ result to each of these three terms to obtain the general result (after arguing similarly for the lower tail). □

Lemma 5. *Suppose $\hat{M}_n$ is a sequence of symmetric positive definite matrices such that uniformly on Γ for some $\delta > 0$,*

$$P_{\boldsymbol{\theta}}\left[0 < \delta \leqslant \lambda_{\min}(\hat{M}_n) \leqslant \lambda_{\max}(\hat{M}_n) \leqslant \frac{1}{\delta}\right] \geqslant 1 - r_n,$$

where $r_n \to 0$ and $\mathrm{E}_{\boldsymbol{\theta}}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 = O(r_n)$ uniformly on Γ. Then

$$\max\{\mathrm{E}_{\boldsymbol{\theta}}\overline{\Phi}(\Psi_{\Sigma}(\hat{\boldsymbol{\Delta}}, \hat{M}_n)) : \boldsymbol{\theta} \in \Gamma\}$$
$$= \max\{\mathrm{E}_{\boldsymbol{\theta}}\overline{\Phi}(\Psi_{\Sigma}(\hat{\boldsymbol{\Delta}}, \hat{M}_n)) : \boldsymbol{\theta} \in \Gamma, c^2 \leqslant \boldsymbol{\Delta}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\Delta} \leqslant f^2\} + O(r_n)$$

for some $f^2 < \infty$.

Proof. It is enough to show that, for any $\varepsilon > 0$, f sufficiently large,

$$\max\{\mathrm{E}_{\boldsymbol{\theta}}\overline{\Phi}(\Psi(\hat{\boldsymbol{\Delta}}, \hat{M})) : \boldsymbol{\theta} \in \Gamma, \boldsymbol{\Delta}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\Delta} \geqslant f^2\} \leqslant \varepsilon + O(r_n).$$

But, evidently, if $\delta \leqslant \lambda_{\min}(\hat{M}) \leqslant \lambda_{\max}(\hat{M}) \leqslant 1/\delta$,

$$\begin{aligned}\Psi_{\Sigma}(\hat{\boldsymbol{\Delta}}, \hat{M}) &\geqslant \frac{\delta^2}{\sqrt{k_2}}\|\hat{\boldsymbol{\Delta}}\| \geqslant \delta^2\sqrt{\frac{k_1}{k_2}}(\hat{\boldsymbol{\Delta}}^{\mathrm{T}}\Sigma^{-1}\hat{\boldsymbol{\Delta}})^{1/2}\\ &\geqslant \delta^2\sqrt{\frac{k_1}{k_2}}(\boldsymbol{\Delta}^{\mathrm{T}}\Sigma^{-1}\boldsymbol{\Delta})^{1/2} + O(r_n)\\ &\geqslant \delta^2\sqrt{\frac{k_1}{k_2}}f + O(r_n).\end{aligned}$$

The lemma follows. □

Lemma 6. *Suppose* M *is a* $(2d+1)$ *diagonal matrix which is symmetric, that is,* $M = \|m_{ab}\|_{p\times p}$,

$$m_{ab} = 0, \qquad |a-b| > d,$$
$$m_{ab} = m_{ba}.$$

Let $\|M\|$ *be the operator norm* $(M : l_2 \to l_2)$

$$\|M\| = \sqrt{\lambda_{\max}(MM^{\mathrm{T}})}$$

Then, for $0 < d < p$,

$$\|M\| \leqslant (2d+1)\|M\|_{\infty} \leqslant (2\sqrt{pd}+1)\|M\|_{\infty},$$

where

$$\|M\|_{\infty} = \max_{a,b}|m_{ab}|.$$

Proof. By symmetry, $\|M\| = \max_{1\leqslant i\leqslant p}|\lambda_i(M)|$, where $\lambda_i(M), \ldots, \lambda_p(M)$ are the eigenvalues of M (real by symmetry). So $\|M\| = \max\{\sup_{\|\mathbf{x}\|=1}\mathbf{x}^{\mathrm{T}}M\mathbf{x}, -\inf_{\|\mathbf{x}\|=1}\mathbf{x}^{\mathrm{T}}M\mathbf{x}\}$. But

$$
\begin{aligned}
|\mathbf{x}^{\mathrm{T}} M \mathbf{x}| &= \left| \sum_{|a-b| \leqslant d} x_a x_b m_{ab} \right| \\
&\leqslant \|M\|_\infty \sum_{k=-d}^{d} \sum_{i=1}^{p} |x_i x_{i+k}| \\
&\leqslant (2d+1)\|M\|_\infty \qquad \text{by Cauchy–Schwarz,} \\
&\leqslant (2\sqrt{pd}+1)\|M\|_\infty .
\end{aligned}
$$

□

Lemma 7. (Kolmogorov's theorem). *Let*

$$\mathcal{F} = \{f : (-\pi, \pi) \to R, \|f^{(r)}\|_\infty \leqslant 1\}.$$

If

$$f(x) = \sum_{k=-\infty}^{\infty} a_k(f) \mathrm{e}^{\mathrm{i}kx},$$

let

$$f_n(x) = \sum_{k=-n}^{n} a_k(f) \mathrm{e}^{\mathrm{i}kx}.$$

Then

$$\sup\{\|f_n - f\|_\infty : f \in \mathcal{F}\} \leqslant C_r \frac{\log n}{n^r}.$$

Proof. See Theorem 1.1 in De Vore and Lorentz (1993, p. 334).

Acknowledgement

We thank Iain M. Johnstone for helpful suggestions, a correction, and discussion.

References

Böttcher, A., Dijksma, A., Langer, H., Dritschel, M., Rovnyak, J. and Kaashoek, M. (1996) *Lectures on Operator Theory and Its Applications.* Providence, RI: American Mathematical Society.

Bradley, T. (2002) On positive spectral density functions. *Bernoulli*, **8**, 175–193.

De Vore, R. and Lorentz, G. (1993) *Constructive Approximation.* Berlin: Springer-Verlag.

Domingos, P. and Pazzani, M. (1997) On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, **29**, 103–130.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Pickard, D. (1995) Wavelet shrinkage: asymptopia? (with discussion). *J. Roy. Statist. Soc. Ser. B*, **57**, 301–369.

Dudoit, S., Fridlyand, J. and Speed, T.P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Statist. Assoc.*, **97**, 77–87.

Greenshtein, E. and Ritov, Y. (2004) Consistency in high dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, **10**, 971–988.

Grenander, U. and Szegö, G. (1984) *Toeplitz Forms and Their Applications*. New York: Chelsea.

Johnstone, I.M. (2002) Function estimation and Gaussian sequence models. Manuscript.

Levina, E. (2002) Statistical issues in texture analysis. PhD thesis, University of California, Berkeley.

Lewis, D.D. (1998) Naive (Bayes) at forty: The independence assumption in information retrieval. In C. Nédellec and C. Rouveirol (eds), *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pp. 4–15. Heidelberg: Springer-Verlag.

Luenberger, D.G. (1984) *Linear and Nonlinear Programming.* Addison-Wesley.

McLachlan, G.J. (1992) *Discriminant Analysis and Statistical Pattern Recognition.* New York: Wiley.

Received January 2003 and revised July 2003