

Some theory for Fisher’s Linear Discriminant  
function, “naive Bayes”, and some alternatives  
when there are many more variables than  
observations.

Peter J. Bickel<sup>1</sup> and Elizaveta Levina<sup>2</sup>

<sup>1</sup> Department of Statistics, UC Berkeley, Berkeley CA 94720-3860,  
bickel@stat.berkeley.edu

<sup>2</sup> Department of Statistics, the University of Michigan,  
Ann Arbor MI 48109-1092, elevina@umich.edu (corresponding author)

June 30, 2003

**Abstract**

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 “colored noise” model and to adapt to a spectrum of convergence rates, which we conjecture to be minimax.

*Keywords:* Fisher’s linear discriminant, Gaussian colored noise, minimax regret, naive Bayes.

*Running Title:* Fisher’s Linear Discriminant and Naive Bayes

# 1 Introduction

It has long been appreciated in machine learning practice (see e.g., Lewis (1998) and 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  $\gg n$ . 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 “colored 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 (2003), 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  $\Sigma$  are known then the optimal classifier is the Bayes rule:

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

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

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

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

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

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  $n_0 \rightarrow \infty$ ,  $n_1 \rightarrow \infty$ , and  $\frac{n_0}{n_1+n_0} \rightarrow \pi$ ,  $0 < \pi < 1$ . If we naturally assume that  $\pi$  is the probability of a new observation belonging to class 0, the rule is modified by replacing 0 in the indicator by  $\log\left(\frac{n_0}{n_1}\right)$ .

**A 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)

$$\delta_F(\mathbf{X}) = \mathbf{1} \left\{ \hat{\boldsymbol{\Delta}}' \hat{\Sigma}^{-1} (\mathbf{X} - \hat{\boldsymbol{\mu}}_\cdot) > 0 \right\}.$$

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

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

and a different discrimination rule:

### Independence Rule (IR)

$$\delta_I(\mathbf{X}) = \mathbf{1} \left\{ \hat{\boldsymbol{\Delta}}' \hat{D}^{-1} (\mathbf{X} - \hat{\boldsymbol{\mu}}_\cdot) > 0 \right\},$$

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

$$\Gamma(c, k, B) = \left\{ (\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \Sigma) : \boldsymbol{\Delta}' \Sigma^{-1} \boldsymbol{\Delta} \geq c^2, k_1 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq k_2, \boldsymbol{\mu}_i \in B, i = 0, 1 \right\},$$

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  $\Sigma$ , and  $B$  is the compact subset of  $l_2$  given by

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

Here,  $a_j \rightarrow \infty$  and  $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots)$ . It is well known that  $B$  is a compact and that (see Pinsker's Theorem in Johnstone (2002)) if  $\Sigma = \text{identity}$ , then for suitable  $r_{jn}$ , depending only on  $\{a_j\}$ , the  $j$ -th component of  $\boldsymbol{\mu}_i$  can be estimated by

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

and then,

$$\max_{\Gamma} E_{\boldsymbol{\theta}} \|\hat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i\|^2 = o(1). \quad (2.3)$$

The condition on eigenvalues guarantees that

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

Then, both  $\Sigma$  and  $\Sigma^{-1}$  are not ill-conditioned. The condition  $\boldsymbol{\Delta}'\Sigma^{-1}\boldsymbol{\Delta} \geq c^2$  guarantees 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  $\boldsymbol{\mu}_0$  to be the same as that under  $\boldsymbol{\mu}_1$ .

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

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

and the worst case posterior error by

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

Further, let

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

be the misclassification error of  $\delta$ , and

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

be the worst case error. For the two rules  $\delta_F$  and  $\delta_I$ , the posterior errors can easily be computed as

$$\begin{aligned} W(\delta_F, \boldsymbol{\theta}) &= \bar{\Phi}(\Psi_{\Sigma}(\hat{\boldsymbol{\Delta}}, \hat{\Sigma})), \\ W(\delta_I, \boldsymbol{\theta}) &= \bar{\Phi}(\Psi_{\Sigma}(\hat{\boldsymbol{\Delta}}, \hat{D})), \end{aligned}$$

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

$$\Psi_{\Sigma}(\mathbf{\Delta}, M) = \frac{\mathbf{\Delta}' M^{-1} \mathbf{\Delta}}{2(\mathbf{\Delta}' M^{-1} \Sigma M^{-1} \mathbf{\Delta})^{1/2}} \quad (2.4)$$

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

$$\bar{W}_{\Gamma}(\delta_F) \rightarrow \bar{\Phi}(c/2),$$

which is the Bayes risk, while  $\bar{W}_{\Gamma}(\delta_I)$  converges to something strictly greater than the Bayes risk. If  $p > n$ ,  $\delta_F$  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  $R^p$  spanned by the eigenvectors  $\hat{\xi}_1, \dots, \hat{\xi}_n$  corresponding to non-zero eigenvalues  $\hat{\lambda}_1, \dots, \hat{\lambda}_n$  of  $\hat{\Sigma}$  and then defining

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

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

**Theorem 1** (a) If  $\frac{p}{n} \rightarrow \infty$ ,  $\bar{W}_{\Gamma}(\delta_F) \rightarrow \frac{1}{2}$ .

(b) If  $\frac{\log p}{n} \rightarrow 0$ , then

$$\limsup_{n \rightarrow \infty} \bar{W}_{\Gamma}(\delta_I) = \bar{\Phi} \left( \frac{\sqrt{K_0}}{1 + K_0} c \right). \quad (2.5)$$

Note that if the matrix has eigenvalues going to 0 or  $\infty$  as  $p \rightarrow \infty$ , then  $K_0 \rightarrow \infty$ , and  $\limsup \bar{W}(\delta_I) = \frac{1}{2}$ , so the worst case of the rule is no better than random guessing. However, if  $\Sigma$  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  $\mathbf{\Delta}$  and  $\Sigma$  are assumed known, the corresponding IR does not lose much in comparison to the Bayes rule. This remains true for the original IR rule under the conditions of Theorem 1 since then  $e_2$  below is the limiting risk of IR. To see that, let

$$\begin{aligned} e_1 &= \bar{\Phi}(\Psi_{\Sigma}(\mathbf{\Delta}, \Sigma)) = \bar{\Phi} \left( \frac{1}{2} (\mathbf{\Delta}' \Sigma^{-1} \mathbf{\Delta})^{1/2} \right) \\ e_2 &= \bar{\Phi}(\Psi_{\Sigma}(\mathbf{\Delta}, D)) = \bar{\Phi} \left( \frac{1}{2} \frac{\mathbf{\Delta}' D^{-1} \mathbf{\Delta}}{(\mathbf{\Delta}' D^{-1} \Sigma D^{-1} \mathbf{\Delta})^{1/2}} \right) \end{aligned}$$

be the errors of the two rules when  $\Delta$ ,  $\Sigma$ , and  $D = \text{diag}(\Sigma)$  are known. If we write  $\Delta_0 = D^{-1/2}\Delta$ , then the efficiency of IR relative to FR is determined by the ratio  $r$  of the arguments of  $\bar{\Phi}$ , where

$$r = \frac{\Psi_{\Sigma}(\Delta, D)}{\Psi_{\Sigma}(\Delta, \Sigma)} = \frac{(\Delta_0' \Delta_0)}{[(\Delta_0' \Sigma_0 \Delta_0)(\Delta_0' \Sigma_0^{-1} \Delta_0)]^{1/2}} \quad (2.6)$$

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  $v$ ,

$$\frac{(v'v)^2}{(v'Qv)(v'Q^{-1}v)} \geq \frac{4aA}{(a+A)^2}$$

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

$$r \geq \frac{2\sqrt{K_0}}{1+K_0} \quad (2.7)$$

and the error of the IR can be bounded by

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

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 IR is comparable to that of 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 leukemia dataset 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  $\Sigma = I$  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  $y_j = \mu_j + n^{-1/2}\varepsilon_j$ ,  $j \geq 1$  where  $\varepsilon_p \equiv (\varepsilon_1, \dots, \varepsilon_p)'$  is Gaussian with mean 0 and variance covariance matrix  $\Sigma_p$ . Let  $\Lambda = \{(\mu_p, \Sigma_p) : \mu \in B_{\mathbf{a},d}, \lambda_{\max}(\Sigma_p) \leq k_2 < \infty\}$  for  $k_2 \geq 1$ . Then there exist  $\hat{\mu} = (\hat{\mu}_1, \hat{\mu}_2, \dots)'$  such that*

$$\max\{E_{\theta}\|\hat{\mu}_p - \mu_p\|^2 : \theta = (\mu, \Sigma) \in \Lambda\} = o(1)$$

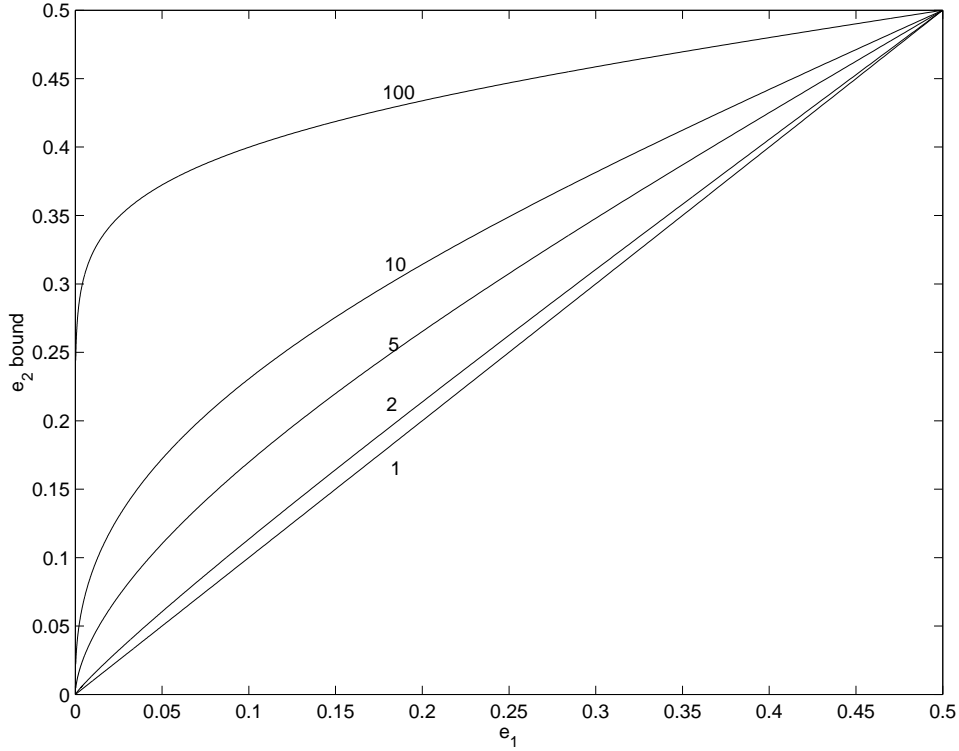


Figure 1: The bound on the risk of the IR as a function of the Bayes risk. The numbers over the curves show the value of  $K_0$ .

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

$$\max\{E_{\boldsymbol{\theta}}\|\hat{\boldsymbol{\mu}}_p - \boldsymbol{\mu}_p\|^2 : \boldsymbol{\theta} \in \Lambda\} \leq k_2 \max\{E_{\boldsymbol{\mu}}\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 : \boldsymbol{\mu} \in B_{a,d}\}.$$

**Proof of Theorem 1:**

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

$$\Psi_I(\hat{\Delta}, \hat{\Sigma}^-) \xrightarrow{P} 0 \tag{2.8}$$

Write, using the spectral theorem,

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

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

$$\Psi_I(\hat{\Delta}, \hat{\Sigma}^-) \leq \sqrt{\sum_{i=1}^n (\hat{\Delta}, \hat{\boldsymbol{\xi}}_i)^2}$$

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

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

$$\max_{1 \leq i \leq p} |\hat{\sigma}_{ii}^{-1} - \sigma_{ii}^{-1}| \xrightarrow{P} 0 \quad (2.9)$$

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

$$\max_{1 \leq i \leq p} |\hat{\sigma}_{ii} - \sigma_{ii}| \xrightarrow{P} 0 \quad (2.10)$$

uniformly on  $\Gamma$ . But by Lemma 4 in Section 6,

$$P \left[ \max_{1 \leq i \leq p} \left| \frac{\hat{\sigma}_{ii}}{\sigma_{ii}} - 1 \right| \geq \varepsilon \right] \leq 2pe^{-nc(\varepsilon)}$$

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

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

$$\max\{W(\delta_I^{\Delta}, \theta) : \theta \in \Gamma\} = \max\{\bar{\Phi}(\Psi_{\Sigma}(\Delta, \hat{D})) : \theta \in \Gamma, \Delta' \Sigma^{-1} \Delta = c^2\} \quad (2.11)$$

On the other hand,

$$\bar{W}(\delta_I, \theta) = E_{\theta} \bar{\Phi}(\Psi_{\Sigma}(\hat{\Delta}, \hat{D}))$$

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

$$\max_{\Gamma} \bar{W}(\delta_I, \theta) = \max\{\bar{W}(\delta_I, \theta) : \theta \in \Gamma, c^2 \leq \Delta' \Sigma^{-1} \Delta \leq f^2 < \infty\} \quad (2.12)$$

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

$$\max\{|\Psi_{\Sigma}(\hat{\Delta}, \hat{D}) - \Psi_{\Sigma}(\Delta, \tilde{D})| : \theta \in \Gamma, c^2 \leq \Delta' \Sigma^{-1} \Delta \leq f^2\} = o_p(1) \quad (2.13)$$

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

$$\begin{aligned} & \min\{\Psi_{\Sigma}(\Delta, D) : \theta \in \Gamma, c^2 \leq \Delta' \Sigma^{-1} \Delta \leq f^2\} \\ &= \min\{\Psi_{\Sigma}(\Delta, D) : \theta \in \Gamma, \Delta' \Sigma^{-1} \Delta \geq c^2\} = \frac{\sqrt{K_0}}{1 + K_0} c \end{aligned} \quad (2.14)$$

To see why this is sufficient to establish result (2.5), note first that by (2.12) we need only consider  $\bar{W}_I$  on the set  $\tilde{\Gamma}$  where  $\theta \in \Gamma$ ,  $c^2 \leq \Delta' \Sigma^{-1} \Delta \leq f^2$ . Next by (2.13) we can replace  $\delta_I$  in  $W$  by  $\delta_I^{\Delta}$  on  $\tilde{\Gamma}$ . Replacing  $\delta_I$  by  $\delta_I^{\Delta}$  in

$W$  implies that this replacement in  $\bar{W}$  is also permitted since  $0 \leq \bar{\Phi} \leq 1$ . Then (2.11) permits us to consider  $W(\delta_I^{\Delta}, \theta)$  and hence  $\bar{W}(\delta_I^{\Delta}, \theta)$  just for  $\theta \in \Gamma$ ,  $\Delta' \Sigma^{-1} \Delta = c^2$ . Again by (2.13), we can replace  $\hat{D}$  in  $\delta_I^{\Delta}$  by  $D$ .

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

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

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

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

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)

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

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

### 3 Connections to Spectral Density

If we think of  $\Sigma_0$  as covariance of a stationary process  $\{\xi_t\}$ , the condition on the eigenvalues of the correlation matrix that assures the good performance of the IR can be related to the spectral density of the corresponding process. Abusing the notation, if we write,

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

then we can think of  $\Sigma_0$  as the correlation matrix for  $p = \infty$ , with correlations for finite  $p$  obtained by taking the first  $p$  rows and columns of  $\Sigma_0$ .

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

$$\sigma(m) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{im\gamma} dF(\gamma)$$

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

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

is the spectral density. Moreover,

$$\begin{aligned} \lambda_{\max}(\Sigma_0) &= \sup_{\nu} f(\nu), \\ \lambda_{\min}(\Sigma_0) &= \inf_{\nu} f(\nu). \end{aligned}$$

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

$$0 < M^{-1} \leq f(\nu) \leq M < \infty \text{ for all } \nu \quad (3.3)$$

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,

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

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

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

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

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

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

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 tri-diagonal 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 rule. 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  $\lambda_1 = \dots = \lambda_{p-1} = 1 - \rho, \lambda_p = 1 + (p - 1)\rho$ , so that  $\lambda_p / \lambda_1 \rightarrow \infty$  as  $p \rightarrow \infty$  and the worst-case error of the IR is also  $1/2$ .

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 “colored” noise model and Bayes consistency and minimax regret

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

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

where  $\boldsymbol{\mu}_i \in l_2$  and  $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots)$  is a sequence of Gaussian variables with mean 0 and  $\text{cov}(\varepsilon_a, \varepsilon_b) = \sigma_{ab}, 1 \leq a \leq b < \infty$ . Let  $\Sigma_p$  be the upper  $p \times p$  corner of  $\Sigma$ , i.e.,  $\|\text{cov}(\varepsilon_i, \varepsilon_j)\|, 1 \leq i, j \leq p$ . We denote  $(\mathbf{a}, \mathbf{b}) \equiv \sum_{i=1}^{\infty} a_i b_i$  for  $\mathbf{a} = (a_1, a_2, \dots)$  and  $\mathbf{b} = (b_1, b_2, \dots)$  and  $\|\mathbf{a}\| = \sum_{i=1}^{\infty} a_i^2$  as usual. Now  $\Sigma \equiv \|\sigma_{ij}\|, 1 \leq i, j < \infty$  is an infinite dimensional matrix. Suppose  $\Sigma$  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 iff  $\sum_{j=1}^{\infty} \sigma_{ij}^2 < \infty$  for all  $i$ . We assume that  $\Sigma$  is bounded and has a bounded inverse, that is, for all  $\mathbf{a} \in l_2$

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

for some  $M$  finite. Such a  $\Sigma$  is a Toeplitz operator, since it satisfies  $\sigma_{ij} = \sigma(|i-j|)$ . If  $O = \|o_{ij}\|_{1 \leq i, j < \infty}$  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,  $\lambda_1(O), \lambda_2(O), \dots$  and

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

It follows that, for  $\Sigma$  as above,

$$\begin{aligned} \|\Sigma\| &= \lambda_{\max}(\Sigma) \leq M \\ \|\Sigma^{-1}\| &= \lambda_{\max}(\Sigma^{-1}) = \lambda_{\min}^{-1}(\Sigma) \leq M \end{aligned} \quad (4.3)$$

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,*

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

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

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

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

$$\|T\| = \sup_x |g_T(x)| \quad (4.4)$$

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

$$\|T^{-1}\| = (\inf_x |g_T(x)|)^{-1} \quad (4.5)$$

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  $\delta_F$  and  $\delta_I$ . We define  $\delta_{I_d}$  as the rule which replaces  $\hat{\Sigma}_p$  by  $\hat{\Sigma}_p^{(d)}$  given below.

$$\begin{aligned}
I_d R : \hat{\Sigma}_p^{(d)} &= \|\tilde{\sigma}_{ab}^{(d)}\|_{p \times p} & (4.6) \\
\tilde{\sigma}_{ab}^{(d)} &= \tilde{\sigma}(b-a), \quad |a-b| \leq d \\
&0 \quad \text{otherwise} \\
\tilde{\sigma}(k) &= (p-k)^{-1} \left\{ \sum_{a=1}^{p-k} \hat{\sigma}_{a, a+k} \right\}
\end{aligned}$$

The rules  $I_d R$  are natural if we assume that  $\Sigma_p$  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  $\Sigma$  over a subset of Toeplitz operators with spectral densities satisfying (3.3) and smoothness restrictions. Let

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

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

$$\Gamma_r = \{\boldsymbol{\theta} : \boldsymbol{\mu}_i \in B_{\mathbf{a}, d}, M^{-1} \leq f_\Sigma \leq M < \infty, (\Sigma^{-1} \boldsymbol{\Delta}, \boldsymbol{\Delta}) \geq c^2, \|f_\Sigma^{(r)}\|_\infty \leq M_r\}$$

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

$$\max\{E_{\boldsymbol{\mu}} \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 : \boldsymbol{\mu} \in B_{\mathbf{a}, d}\} \asymp n^{-\alpha}, \quad \alpha < 1 \quad (4.7)$$

and let

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

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

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

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

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

**Proof of Theorem 2:**

We will write  $\Delta_p$  to signify the first  $p$  coordinates of  $\Delta \in l_2$ . We claim that, for all  $\theta \in \Gamma_r$  with  $c^2 \leq \Delta_p' \Sigma_p^{-1} \Delta_p \leq f^2$ ,

$$\begin{aligned} & |\Psi_{\Sigma_p}(\Delta_p^*, \Sigma_p^*) - \Psi_{\Sigma_p}(\Delta_p, \Sigma_p)| \\ & \leq C\{\|\Delta_p^* - \Delta_p\|^2 + \|\Sigma_p^* - \Sigma_p\|^2\} \end{aligned} \quad (4.10)$$

for all  $\|\Delta_p^* - \Delta_p\| \leq \delta_1$ ,  $\|\Sigma_p^* - \Sigma_p\| \leq \delta_2$ , for  $\delta_1, \delta_2$  small enough and  $C$  depending on  $\Gamma_r$  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)  $D\Psi_{\Sigma_p}(\Delta_p, \Sigma_p) = 0$
- b)  $\sup\{|D^2\Psi_{\Sigma_p}(\Delta_p^*, \Sigma_p^*)| : \|\Delta_p^* - \Delta_p\| \leq \delta_1, \|\Sigma_p^* - \Sigma_p\| \leq \delta_2\} \leq C < \infty$  for all  $\theta$  as above.

Until we need them again, we shall drop the  $p$  subscripts. To show a), we expand the numerator of  $\Psi$  as

$$\begin{aligned} & (\Delta + \lambda_1 \mathbf{e})'(\Sigma^{-1} + \lambda_2 E)(\Delta + \lambda_1 \mathbf{e}) \\ & = \Delta' \Sigma^{-1} \Delta + 2\lambda_1 \mathbf{e}' \Sigma^{-1} \Delta + \lambda_2 \Delta' E \Delta + O(\lambda_1 + \lambda_2)^2, \end{aligned} \quad (4.11)$$

and the denominator of  $\Psi$  as

$$\begin{aligned} & (\Delta + \lambda_1 \mathbf{e})'(\Sigma^{-1} + \lambda_2 E)\Sigma(\Sigma^{-1} + \lambda_2 E)(\Delta + \lambda_1 \mathbf{e}) \\ & = (\Delta' \Sigma^{-1} \Delta + 2\lambda_2 \Delta' E \Delta + 2\lambda_1 \mathbf{e}' \Sigma^{-1} \Delta) + O(\lambda_1 + \lambda_2)^2 \\ & = (\Delta' \Sigma^{-1} \Delta)^{\frac{1}{2}}(1 + \lambda_1 \mathbf{e}' \Sigma^{-1} \Delta + \lambda_2 \Delta' E \Delta) + O(\lambda_1 + \lambda_2)^2. \end{aligned} \quad (4.12)$$

Hence,

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

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  $|\lambda_1| \leq \delta_1$ ,  $|\lambda_2| \leq \delta_2$ ,  $\|\mathbf{e}\| \leq 1$ ,  $\|E\| \leq 1$ . The upper bounds are straightforward. For instance,

$$\begin{aligned} [\Delta^*]' E \Delta^* & \leq \|\Delta^*\|^2 \leq 2(\|\Delta\|^2 + \delta_1^2) \\ & \leq 2(k_2 f^2 + \delta_1^2) \end{aligned}$$

since  $\|\Delta\|^2/k_2 \leq \Delta' \Sigma^{-1} \Delta \leq f^2$ .

On the other hand,

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

Hence,

$$\begin{aligned} [\mathbf{\Delta}^*]'[\Sigma^*]^{-1}\Sigma[\Sigma^*]^{-1}[\mathbf{\Delta}^*] &\geq \frac{k_1}{(k_2 + \delta_2)^2} \|\mathbf{\Delta}^*\|^2 \\ &\geq \frac{k_1}{(k_2 + \delta_2)^2} (\|\mathbf{\Delta}\| - \delta_1)^2 \geq \frac{k_1}{(k_2 + \delta_2)^2} (c\sqrt{k_1} - \delta_1)^2 \end{aligned} \quad (4.13)$$

which is bounded away from 0 for  $\delta_1$  small. Claim b) follows.

From (4.10), we see that

$$\begin{aligned} &|E_{\boldsymbol{\theta}} \bar{\Phi}(\Psi_{\Sigma_p}(\hat{\mathbf{\Delta}}_p, \hat{\Sigma}_p^{(d)})) - E_{\boldsymbol{\theta}} \bar{\Phi}(\Psi_{\Sigma_p}(\mathbf{\Delta}_p, \Sigma_p^{(d)}))| \\ &\leq C(E_{\boldsymbol{\theta}} \|\hat{\mathbf{\Delta}}_p - \mathbf{\Delta}_p\|^2 \mathbf{1}(\|\hat{\mathbf{\Delta}}_p - \mathbf{\Delta}_p\| \leq \delta_1) + E_{\boldsymbol{\theta}} \|\hat{\Sigma}_p^{(d)} - \Sigma_p\|^2 \mathbf{1}(\|\hat{\Sigma}_p - \Sigma_p\| \leq \delta_2)) \\ &+ P[\|\hat{\mathbf{\Delta}}_p - \mathbf{\Delta}_p\| \geq \delta_1] + P[\|\hat{\Sigma}_p^{(d)} - \Sigma_p\| \geq \delta_2] \end{aligned} \quad (4.14)$$

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

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

where  $\bar{X}_j = \frac{1}{n} \sum_{k=1}^n X_{kj}$ . We use Lemma 4 to bound, for  $a \geq 1$ ,  $P[|\tilde{\sigma}(a) - \sigma(a)| \geq \nu]$ . We can apply Lemma 4 since  $(X_{k1}, \dots, X_{kp})$  are i.i.d.  $N(\mathbf{0}, \Sigma_p)$  and  $\sqrt{n}(\bar{X}_1, \dots, \bar{X}_p)$  are  $N(\mathbf{0}, \Sigma_p)$  as well.

By Lemma 4,

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

Suppose  $d \leq \frac{p}{2}$ . Then, for some  $\varepsilon > 0$ ,  $A < \infty$ ,  $\delta > 0$ ,  $\delta \geq \nu \geq \frac{A}{\sqrt{np}}$ ,

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

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

Let,

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

By Lemma 6,

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

Therefore, by (4.16)

$$\begin{aligned} &E_{\boldsymbol{\theta}} \|\hat{\Sigma}_p^{(d)} - \Sigma_p^{(d)}\|^2 \mathbf{1}(\|\hat{\Sigma}_p^{(d)} - \Sigma_p^{(d)}\| \leq \delta) \\ &\leq 4pd \int_0^\delta \nu P[V_d \geq \nu] d\nu \\ &\leq 4pd \left( \frac{A^2}{2np} + \int_{\frac{A}{\sqrt{np}}}^\infty \nu e^{-\varepsilon np \nu^2} d\nu \right) \\ &\leq B_1 \frac{d}{n} \end{aligned} \quad (4.17)$$

for some universal  $B_1$ . On the other hand,

$$P[\|\hat{\Sigma}_p^{(d)} - \Sigma_p^{(d)}\| \geq \delta] \leq B_2 e^{-n\varepsilon} \leq B_2 \frac{d}{n} \quad (4.18)$$

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

$$\begin{aligned} \sup\{|E_{\boldsymbol{\theta}}(\bar{\Phi}(\Psi_{\Sigma_p}(\hat{\boldsymbol{\Delta}}_p, \hat{\Sigma}_p^{(d)}))) - \bar{\Phi}(\Psi_{\Sigma_p}(\boldsymbol{\Delta}_p, \Sigma_p^{(d)}))| : \\ c^2 \leq \boldsymbol{\Delta}'\Sigma^{-1}\boldsymbol{\Delta} \leq f^2, \boldsymbol{\theta} \in \Gamma_r\} \leq C \max\{\frac{d}{n}, n^{-\alpha}\} \end{aligned} \quad (4.19)$$

Now we appeal to Lemma 7, which yields that,

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

$$\|\Sigma_p - \Sigma\| \leq \frac{B \log p}{p^r} \quad (4.21)$$

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

$$\begin{aligned} \bar{W}_{\Gamma_r}(\delta_{Id}) &= \sup\{E_{\boldsymbol{\theta}}\bar{\Phi}\Psi_{\Sigma_p}(\hat{\boldsymbol{\Delta}}_p, \hat{\Sigma}_p^{(d)}) : \boldsymbol{\theta} \in \Gamma_r, c^2 \leq \boldsymbol{\Delta}'_p \Sigma_p^{-1} \boldsymbol{\Delta}_p \leq f^2\} \\ &\quad + O(P[\|\hat{\Sigma}_p^{(d)} - \Sigma_p^{(d)}\| \geq \delta]) \end{aligned} \quad (4.22)$$

Putting (4.20) – (4.22) together, we obtain that if  $d \leq \frac{p}{2}$ ,

$$\begin{aligned} \bar{W}_{\Gamma_r}(\delta_{Id}) &= \sup\left\{\bar{\Phi}\left(\frac{1}{2}(\boldsymbol{\Delta}'_p \Sigma_p^{-1} \boldsymbol{\Delta}_p)^{\frac{1}{2}}\right) : \boldsymbol{\Delta}'_p \Sigma_p^{-1} \boldsymbol{\Delta}_p \geq c^2\right\} \\ &\quad + O\left(\max\left\{\frac{d}{n}, n^{-\alpha}, \frac{(\log d)^{2r}}{d^{2r}}\right\}\right) \\ &= \bar{\Phi}\left(\frac{c}{2}\right) + O(n^{-\gamma} \log n) \end{aligned} \quad (4.23)$$

by taking  $p$  sufficiently large,  $d = \lceil n(\log n)^{2r} \rceil^{\frac{1}{2r+1}}$ . The theorem follows.  $\square$

## 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 0 and should be estimated that way. 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 as in 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 (2003) propose to take advantage of sparsity of  $\Sigma^{-1}\mathbf{\Delta}$  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  $\Sigma$  general:** Note that in the case  $\Sigma = I$  (Johnstone (2002))

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

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

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

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

For arbitrary  $\Sigma_p$ , let  $[\tau_{ij}] = \Sigma_p^{1/2}$  and estimate  $\mu_i$  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  $\delta_j$  are iid  $N(0, \frac{1}{n})$ . Thus,

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

Note that

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

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

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

$$\begin{aligned} M(\nu) &= \max \left\{ \sum_{j=1}^p [\log(1 - 2\gamma_j s) + 2s\gamma_j(\nu + 1)] : 0 \leq s \leq \frac{1}{2\lambda_2} \right\} \\ &\geq \min \left\{ \frac{p\lambda_1^2 \epsilon^2 \nu^2}{2\lambda_2^3}, \frac{p(1 - \epsilon)\nu}{2\lambda_2} \right\} \end{aligned}$$

**Proof:** Note that if  $0 \leq x \leq 1 - \epsilon$ ,

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

Therefore,

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

Substituting

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

for  $\nu \leq \frac{1 - \epsilon}{\epsilon^2}$ , we obtain

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

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

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

for  $\nu \geq 2\lambda_2(\log \frac{1}{\epsilon})(1 - \epsilon)^{-1}$  by taking  $s = (1 - \epsilon)/(2\lambda_2)$ . The lemma follows by taking  $\epsilon$  so that  $2\lambda_2(\log \frac{1}{\epsilon})(1 - \epsilon)^{-2} \leq \epsilon^{-2}$ .  $\square$

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

Then

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

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

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

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  $\Sigma$  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  $\lambda_1 \leq \gamma_1 \leq \dots \leq \gamma_p \leq \lambda_2$  are the eigenvalues of  $\Sigma$  and the  $U_{ij}$  are independent  $N(0, 1)$ .

By Markov's inequality,

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

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,

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

Further write the first term as the sum of two terms

$$\begin{aligned} & \sum_{i=1}^n \sum_{j \in I_1} \{(Z_{ij} + Z_{i(j+t)})^2 - (1 + \sigma(t))\} \\ & + \sum_{i=1}^n \sum_{j \in I_2} \{(Z_{ij} + Z_{i(j+t)})^2 - (1 + \sigma(t))\} \end{aligned}$$

where  $j \in I_1 \Leftrightarrow j+t \in I_2$  and the cardinalities of  $I_1$  and  $I_2$  are approximately  $\frac{p-t}{2}$ , i.e., 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  $l = \frac{p-t}{2}$  or  $\frac{p-t+1}{2}$  and  $\frac{p-t-1}{2}$  and  $\gamma(0) \equiv 1$ .

The maximal eigenvalue of  $\Gamma$  is given by

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

since  $(1 + \sigma(t)) = \text{Var}(\frac{Z_{11} + Z_{1t}}{\sqrt{2}}) \geq \lambda_1$ .

We get a new bound,

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

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

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

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

where  $r_n \rightarrow 0$  and  $E_{\boldsymbol{\theta}} \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 = O(r_n)$  uniformly on  $\Gamma$ .

Then

$$\begin{aligned} & \max\{E_{\boldsymbol{\theta}}\bar{\Phi}\Psi_{\Sigma}(\hat{\Delta}, \hat{M}_n) : \boldsymbol{\theta} \in \Gamma\} \\ &= \max\{E_{\boldsymbol{\theta}}\bar{\Phi}\Psi_{\Sigma}(\hat{\Delta}, \hat{M}_n) : \boldsymbol{\theta} \in \Gamma, c^2 \leq \Delta'\Sigma^{-1}\Delta \leq f^2\} + O(r_n) \end{aligned}$$

for some  $f^2 < \infty$ .

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

$$\max\{E_{\boldsymbol{\theta}}\bar{\Phi}\Psi_{\Sigma}(\hat{\Delta}, \hat{M}) : \boldsymbol{\theta} \in \Gamma, \Delta'\Sigma^{-1}\Delta \geq f^2\} \leq \varepsilon + O(r_n)$$

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

$$\begin{aligned} \Psi_{\Sigma}(\hat{\Delta}, \hat{M}) &\geq \frac{\delta^2}{\sqrt{k_2}} \|\hat{\Delta}\| \geq \delta^2 \sqrt{\frac{k_1}{k_2}} (\hat{\Delta}'\Sigma^{-1}\hat{\Delta})^{\frac{1}{2}} \\ &\geq \delta^2 \sqrt{\frac{k_1}{k_2}} (\Delta'\Sigma^{-1}\Delta)^{\frac{1}{2}} + O(r_n) \\ &\geq \delta^2 \sqrt{\frac{k_1}{k_2}} f + O(r_n) \end{aligned}$$

The lemma follows.  $\square$

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

$$\begin{aligned} m_{ab} &= 0, & |a-b| > d \\ m_{ab} &= m_{ba} \end{aligned}$$

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

$$\|M\| = \sqrt{\lambda_{\max}(MM')}$$

Then, for  $0 < d < p$

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

where

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

**Proof:** By symmetry,  $\|M\| = \max_{1 \leq i \leq p} |\lambda_i(M)|$ , where  $\lambda_1(M), \dots, \lambda_p(M)$  are the eigenvalues of  $M$  (real by symmetry). So  $\|M\| = \max\{\sup_{\|\mathbf{x}\|=1} \mathbf{x}'M\mathbf{x}, -\inf_{\|\mathbf{x}\|=1} \mathbf{x}'M\mathbf{x}\}$ . But

$$\begin{aligned} |\mathbf{x}'M\mathbf{x}| &= \left| \sum_{|a-b| \leq d} x_a x_b m_{ab} \right| \\ &\leq \|M\|_{\infty} \sum_{k=-d}^d \sum_{i=1}^p |x_i x_{i+k}| \\ &\leq (2d+1) \|M\|_{\infty} \quad \text{by Cauchy-Schwartz,} \\ &\leq (2\sqrt{pd}+1) \|M\|_{\infty} \end{aligned}$$

□

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

$$\mathcal{F} = \{f : (-\pi, \pi) \rightarrow \mathbb{R}, \|f^{(r)}\|_\infty \leq 1\}$$

*If,*

$$f(x) = \sum_{k=-\infty}^{\infty} a_k(f) e^{ikx}$$

*let*

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

*Then,*

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

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

## Acknowledgment

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

## References

- Böttcher, A., Dijkma, A., Langer, H., Dritschel, M., Rovnyak, J. and Kaashoek, M. (1996). *Lectures on operator theory and its applications*, American Mathematical Society, Providence, Rhode Island.
- Bradley, T. (2002). On positive spectral density functions, *Bernoulli* **8**(2): 175–193.
- De Vore, R. and Lorentz, G. (1993). *Constructive Approximation*, Springer-Verlag, Berlin, New York.
- 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, *JASA* **97**(457): 77–87.
- Greenshtein, E. and Ritov, Y. (2003). Consistency in high dimensional linear predictor selection and the virtue of over-parametrization, *Bernoulli* (To appear).
- Grenander, U. and Szegö, G. (1984). *Toeplitz Forms and Their Applications*, Chelsea Publishing Company, New York.
- Johnstone, I. M. (2002). *Function estimation and Gaussian sequence models*, Draft of book in preparation.
- Levina, E. (2002). *Statistical Issues in Texture Analysis*, PhD thesis, UC 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*, Springer Verlag, Heidelberg, DE, pp. 4–15.
- Luenberger, D. G. (1984). *Linear and Nonlinear Programming*, Addison-Wesley.
- McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York.