Classification of sparse high-dimensional vectors

By Yuri I. Ingster1, Christophe Pouet2,* and Alexandre B. Tsybakov3

1St. Petersburg State Electrotechnical University, 5 Prof. Popov Street, 197376 St. Petersburg, Russia
2LATP, University of Provence, 39 rue F. Joliot-Curie, 13453 Marseille Cedex 13, France
3Laboratoire de Statistique, CREST, Timbre J340, 3 avenue Pierre Larousse, 92240 Malakoff Cedex, France, and LPMA, University of Paris 6, 4 Place Jussieu, 75252 Paris Cedex 05, France

We study the problem of classification of $d$-dimensional vectors into two classes (one of which is ‘pure noise’) based on a training sample of size $m$. The main specific feature is that the dimension $d$ can be very large. We suppose that the difference between the distribution of the population and that of the noise is only in a shift, which is a sparse vector. For Gaussian noise, fixed sample size $m$, and dimension $d$ that tends to infinity, we obtain the sharp classification boundary, i.e. the necessary and sufficient conditions for the possibility of successful classification. We propose classifiers attaining this boundary. We also give extensions of the result to the case where the sample size $m$ depends on $d$ and satisfies the condition $(\log m)/\log d \to \gamma$, $0 \leq \gamma < 1$, and to the case of non-Gaussian noise satisfying the Cramér condition.

Keywords: Bayes risk; classification boundary; high-dimensional data; optimal classifier; sparse vectors

1. Introduction

Over the last decade, high-dimensional data have become a common reality in many fields, for example, in microarray analysis, astronomy, information technology, etc. Often the task is to perform a prediction or to classify a $d$-dimensional observation vector using the information from a training sample of vectors of the same dimension. The main difficulty is that the dimension $d$ can be very large. For example, in gene microarray data, $d$ is typically of the order $10^5$, so that, speaking in mathematical terms, it is natural to consider the asymptotics in the dimension, i.e. as $d \to \infty$. This situation is highly non-standard, in the sense that it is not covered by the classical statistical techniques.

In this paper, we study the two-class classification problem in such a framework. When the dimension $d$ is large, there is no hope of constructing a successful classification method for vectors of general form. However, we shall show that it becomes possible if the underlying high-dimensional vectors have

*Author for correspondence (pouet@cmi.univ-mrs.fr).

One contribution of 11 to a Theme Issue ‘Statistical challenges of high-dimensional data’.
some simple structure. Below, we shall assume that they are sparse, i.e. contain only a small fraction of non-zero components described by the sparsity index $\beta$. This assumption is natural in many applications. We shall present efficient classification methods under such a sparsity assumption. What is more, we shall establish the classification boundary, i.e. the necessary and sufficient conditions for the possibility of successful classification. These conditions characterize the classification problem and not a particular method. They can be helpful for the practitioner who needs to design experiments. Modern experiments, e.g. in bioinformatics, are often resource intensive (time, raw material, cost). Our results on the classification boundary can give guidelines to decide whether or not an experiment is worth performing. Indeed, it turns out that the regions where the practitioner can or cannot expect good performance in classification are determined by two parameters of the problem, namely, the sparsity index $\beta$ and a kind of signal-to-noise ratio. The most striking result here is that the regions are completely described through very simple conditions on these two parameters. Knowing in advance that the parameters of the experiment are correct regarding our theoretical results can help the practitioner to conserve resources. On the other hand, our theory would give some hints about how to improve the experimental design. Another striking result is that successful classification can be expected even if the coordinates of the signal cannot be accurately estimated. For example, in cancer classification using gene microarray data, our results show that, for a typical set-up with $m = 10$ patients and more than $d = 20,000$ genes (see Yeoh et al. 2002; Ross et al. 2003, 2004), the signal-to-noise ratio can be smaller than 1 even when very few coordinates are informative. This point is very important and can lead one to discard a broadly used methodology that consists of selecting genes prior to classification in the diagnosis problem. Indeed, our results show that from the diagnosis point of view the practitioner should use the whole range of data. This also suggests taking a new look at the previously collected data that were uninformative with the traditional methodology.

(a) Model and problem

Assume that we have an independent and identically distributed (i.i.d.) sample $\mathbf{Y} = (Y_1, \ldots, Y_m)$ from a population with probability distribution $P_Y$ on $\mathbb{R}^d$. Here,

$$Y_j = (Y^1_j, \ldots, Y^d_j),$$

where $Y^k_j$ are the components of $Y_j$. In this paper we assume that

$$Y^k_j = u_k + \eta^k_j,$$  \hspace{1cm} (1.1)

where $u = (u_1, \ldots, u_d)$ is a deterministic mean vector and the errors $\eta^1_j, \ldots, \eta^d_j$ are (unless other conditions are explicitly mentioned) jointly i.i.d. zero-mean random variables with probability density $f$ on $\mathbb{R}$. Let $P_\eta$ denote the probability distribution of $(\eta^1_1, \ldots, \eta^d_1)$.

We consider the classification problem when the dimension of the observations $d$ is very large (tends to $+\infty$). Assume that we observe a random vector $Z = (Z^1, \ldots, Z^d)$ independent of $\mathbf{Y}$ and we know that the distribution of $Z$ is either...
$P_\eta$ or $P_Y$. Our aim is to classify $Z$, i.e. to decide whether $Z$ comes from the population with distribution $P_Y$ or from the ‘pure noise’ population $P_\eta$.

Distinguishing between $P_\eta$ and $P_Y$ presents a difficulty only when the vector $u$ is close to 0. A particular type of closeness for large $d$ can be characterized by the sparsity assumption (Ingster & Suslina 2002a; Donoho & Jin 2004) that we shall adopt in this paper. Similar to Ingster & Suslina (2002a) and Donoho & Jin (2004), we introduce the following set of sparse vectors in $\mathbb{R}^d$ characterized by a positive number $a_d$ and a sparsity index $\beta \in (0,1)$:

$$U_{\beta,a_d} = \left\{ u = (u_1, \ldots, u_d) : u_k = a_d \varepsilon_k, \varepsilon_k \in [0,1], \sum_{k=1}^d \varepsilon_k \leq C d^{1-\beta} \right\}.$$  

Here, $0 < c < C < +\infty$ are two constants that are supposed to be fixed throughout the paper. The value $p = d^{-\beta}$ can be interpreted as the ‘probability’ of occurrence of non-zero components in vector $u$. A similar model was also considered recently by Hall et al. (2008).

In this paper, we establish the classification boundary, i.e. we specify the necessary and sufficient conditions on $\beta$ and $a_d$ such that successful classification is possible. Let us first define the notion of successful classification. We shall need some notation. Let $\psi$ be a decision rule, i.e. a measurable function of $Y,Z$ with values in $[0,1]$. If $\psi = 0$, we allocate $Z$ to the $P_\eta$ population, whereas for $\psi = 1$ we allocate $Z$ to the $P_Y$-population. The rules $\psi$ taking intermediate values in $(0,1)$ can be interpreted as randomized decision rules. Let $P^{(u)}_{H_0}$ and $P^{(u)}_{H_1}$ denote the joint probability distributions of $Y,Z$ when $Z$ is only noise and $Z \sim P_Y$, respectively, and let $E^{(u)}_{H_0}$, $\text{Var}^{(u)}_{H_0}$ and $E^{(u)}_{H_1}$, $\text{Var}^{(u)}_{H_1}$ denote the corresponding expectation and variance operators. We shall also denote by $P^{(u)}$ the distribution of $Y$ and by $E^{(u)}$, $\text{Var}^{(u)}$ the corresponding expectation and variance operators. Consider the Bayes risk,

$$R_B(\psi) = \pi E^{(u)}_{H_0}(\psi) + (1 - \pi) E^{(u)}_{H_1}(1 - \psi),$$

where $0 < \pi < 1$ is a prior probability of the noise population, and the maximum risk,

$$R_M(\psi) = \max\{E^{(u)}_{H_0}(\psi), E^{(u)}_{H_1}(1 - \psi)\}.$$

Let $R(\psi)$ be either the Bayes risk $R_B(\psi)$ or the maximum risk $R_M(\psi)$.

We shall say that successful classification is possible if $\beta$ and $a_d$ are such that

$$\liminf_{d \to +\infty} \sup_{\psi \in U_{\beta,a_d}} R(\psi) = 0,$$

for $R = R_M$ and $R = R_B$ with any fixed $0 < \pi < 1$. Conversely, we shall say that successful classification is impossible if $\beta$ and $a_d$ are such that

$$\liminf_{d \to +\infty} \inf_{\psi \in U_{\beta,a_d}} R(\psi) = R_{\text{max}},$$

where $R_{\text{max}} = 1/2$ for $R = R_M$ and $R_{\text{max}} = \min(\pi,1-\pi)$ for $R = R_B$ with $0 < \pi < 1$.

We call equation (1.2) the upper bound of classification and equation (1.3) the lower bound of classification. The lower bound equation (1.3) for the maximum risk $R = R_M$ is interpreted as the fact that no decision rule is better (in a minimax
sense) than a simple random guess. For the Bayes risk $R_B$, the lower bound equation (1.3) is attained at the degenerate decision rule that does not depend on the observations: $\psi \equiv 0$ if $\pi > 1/2$ or $\psi \equiv 1$ if $\pi \leq 1/2$.

The condition on $(\beta, a_d)$ corresponding to the passage from equation (1.2) to (1.3) is called the *classification boundary*. We shall say that a classifier $\psi = \psi_d$ is *asymptotically optimal* (or that $\psi$ *attains the classification boundary*) if, for all $\beta$ and $a_d$ such that successful classification is possible, we have

$$\lim_{d \to +\infty} \sup_{u \in U_{\beta, a_d}} R(\psi) = 0,$$

(1.4)

where $R = R_M$ or $R = R_B$ with any fixed $0 < \pi < 1$.

(b) *Summary of the main results*

According to the value of $\beta$, we shall distinguish between *moderately sparse vectors* and *highly sparse vectors*. This division depends on the relationship between $m$ and $d$. For $m$ not too large, i.e. when $\log m = o(\log d)$, moderately sparse vectors correspond to $\beta \in (0, 1/2]$ and highly sparse vectors to $\beta \in (1/2, 1)$. For large $m$, i.e. when $\log m \sim \gamma \log d$, $\gamma \in (0, 1)$, moderate sparsity corresponds to $\beta \in (0, (1 - \gamma)/2]$ and high sparsity to $\beta \in ((1 - \gamma)/2, 1 - \gamma)$.

The classification boundary for moderately sparse vectors is obtained in a relatively simple way (cf. §2). It is of the form

$$R_d \triangleq d^{1/2 - \beta} a_d \asymp 1.$$

This means that successful classification is possible if $R_d \to +\infty$, and it is impossible if $R_d \to 0$ as $d \to +\infty$. The result is valid both for $\beta \in (0, 1/2]$ and $m \geq 1$ fixed or for $m$ depending on $d$, such that $\log m \sim \gamma \log d$, $\gamma \in (0, 1)$ as $d \to +\infty$, and for $\beta \in (0, (1 - \gamma)/2]$. Moreover, equation (1.5) holds under weak assumptions on the noise. In particular, for the upper bound of classification we only need to assume that the noise has mean 0 and finite second moment (cf. §2). The lower bound is proved under a mild regularity condition on the density $f$ of the noise.

The case of highly sparse vectors is more involved. We establish the classification boundary for the following scenarios:

(A) $m \geq 1$ is a fixed integer, and the noise density $f$ is Gaussian $\mathcal{N}(0, \sigma^2)$ with known or unknown $\sigma > 0$.

(B) $m \to +\infty$ as $d \to +\infty$, $\log m = o(\log d)$, and $f$ is Gaussian $\mathcal{N}(0, \sigma^2)$ with known or unknown $\sigma > 0$.

(C) $\log m \sim \gamma \log d$, $\gamma \in (0, 1)$, and $f$ is Gaussian $\mathcal{N}(0, \sigma^2)$ with known or unknown $\sigma > 0$.

The upper bounds are extended to the following additional scenario:

(D) $m \to +\infty$ as $d \to +\infty$, $(\log m)/\log d \to \gamma$, $0 \leq \gamma < 1$, $m/\log d \to +\infty$, and the noise satisfies the Cramér condition.

The conditions on the noise in (A)–(D) are crucial and, as we shall see later, they suggest that a special dependence of $a_d$ on $d$ and $m$ of the form

*Phil. Trans. R. Soc. A* (2009)
Sparse high-dimensional vector classes

$\alpha_d \approx \sqrt{(\log d)/m}$ is meaningful in the highly sparse case. More specifically, we take

$$a_d = s \sigma \sqrt{\log d} \quad \text{and} \quad x_1 = s \sqrt{m + 1}, \quad (1.6)$$

where $x_1 > 0$ is fixed. The classification boundary in (A), (B) and (D) is then expressed by the following condition on $\beta$, $s$ and $m$:

$$x_1 = \phi(\beta), \quad (1.7)$$

where

$$\phi(\beta) = \begin{cases} \phi_1(\beta) & \text{if } \frac{1}{2} < \beta \leq \frac{3}{4}, \\ \phi_2(\beta) & \text{if } \frac{3}{4} < \beta < 1, \end{cases} \quad (1.8)$$

with

$$\phi_1(\beta) = \sqrt{2\beta - 1} \quad \text{and} \quad \phi_2(\beta) = \sqrt{2(1 - \sqrt{1 - \beta})}. \quad (1.9)$$

In other words, successful classification is possible if $x_1 \geq \phi(\beta) + \delta$, and it is impossible if $x_1 \leq \phi(\beta) - \delta$, for any $\delta > 0$ and $d$ large enough. This classification boundary is also extended to the case where $x_1$ depends on $d$ but stays bounded.

For scenario (C), let $a_d = \sigma x \sqrt{(\log d)/m}$ with fixed $x > 0$. We show that in this framework successful classification is impossible if $\beta > 1 - \gamma$ (cf. remark 1 in §2), and therefore we are interested in $\beta \in ((1 - \gamma)/2, 1 - \gamma)$. Set $\beta^* = \beta/(1 - \gamma) \in (1/2, 1)$ and $x^* = x/\sqrt{1 - \gamma}$. Then the classification boundary is of the form

$$x^* = \phi(\beta^*), \quad (1.10)$$

for the function $\phi(\beta)$ defined above. This means that successful classification is possible if $x^* > \phi(\beta^*)$, and it is impossible if $x^* < \phi(\beta^*)$ as $d \to \infty$.

For $m = 0$ (i.e. when there is no sample $Y$ and only $Z$ is observed), the problem that we consider here reduces to the problem of signal detection with growing dimension $d$ (Ingster 1997; Ingster & Suslina 2001a,b, 2002a,b; Jin 2003, 2004; Donoho & Jin 2004; Jager & Wellner 2007), and our classification boundary coincides with the detection boundary established by Ingster (1997). Sharp asymptotics in the detection problem were studied by Ingster (1997) (see also Ingster & Suslina 2002a, ch. 8) for known $a_d$ or $\beta$. The adaptive version of the problem (this corresponds to unknown $a_d$ and $\beta$) was studied by Ingster & Suslina (2001a,b). Various procedures to attain the detection boundary were proposed by Ingster & Suslina (2002b), Donoho & Jin (2004) and Jager & Wellner (2007). Ingster & Suslina (2002b) introduced a method to attain the detection boundary based on the combination of three different procedures for the zones $\beta \in (0, 1/2]$, $\beta \in (1/2, 3/4]$ and $\beta \in (3/4, 1)$. Later, Donoho & Jin (2004) showed that a test based on the higher criticism (HC) statistic attains the detection boundary simultaneously for these zones. More recently, Jager & Wellner (2007) proved that the same is true for a large class of statistics including the HC statistic. Hall & Jin (2009) have extended those results to the case of correlated noise. Haupt et al. (2008) showed that the rates of detection can be improved when using adaptive multiple sampling schemes.

Inference on the fraction of non-zero components in the model with $m = 0$ has been studied by Cai et al. (2007). They derived minimax rates of convergence for estimators of the fraction and the corresponding confidence sets.

Phil. Trans. R. Soc. A (2009)
Hall et al. (2008) deal with the same classification model as the one that we consider here, but study a problem that is different from ours. They analyse the conditions under which some simple (e.g. minimum-distance) classifiers $\psi$ satisfy
\[
\lim_{d \to +\infty} E_{H_0}^{(u)}(\psi) = 0. \tag{1.11}
\]
Hall et al. (2008) conclude that, for minimum-distance classifiers, equation (1.11) holds if and only if $0 < \beta < 1/2$. This implies that such classifiers cannot be optimal for $1/2 \leq \beta < 1$. They also derive equation (1.11) for some other classifiers in the case $m = 1$.

The results of this paper and their extensions to the multi-class setting were summarized by Pouet (2008) and presented at the meeting ‘Rencontres de Statistique Mathématique’ (Luminy, 16–21 December 2008) and at the Oberwolfach meeting ‘Sparse Recovery Problems in High Dimensions: Statistical Inference and Learning Theory’ (15–21 March 2009). A detailed version with full proofs is given by Ingster et al. (2009). In work parallel to ours, Donoho & Jin (2008, in press) and J. Jin (2009, unpublished data) independently and contemporaneously have analysed a setting less general than the present one. They did not consider a minimax framework, but rather demonstrated that the HC methodology can be successfully extended to the classification problem. Donoho & Jin (in press) showed that, for a special case of scenario (B), the ‘ideal’ (i.e. depending on the unknown $a_d$ and $\beta$) HC statistic attains the same upper bound of classification that we prove below. Together with our lower bound, this implies that the ‘ideal’ HC statistic is asymptotically optimal, in the sense defined above, for scenario (B). Donoho and Jin announce that similar results for the HC statistic in scenarios (A) and (C) will appear in their work in preparation.

This paper is organized as follows. Section 2 contains some preliminary remarks. In §3, we present the classification boundary and asymptotically optimal classifier for moderately sparse vectors under rather general conditions on the noise. In §4, we give the classification boundary and asymptotically optimal classifiers for highly sparse vectors under scenarios (A), (B) and (C). Section 5 provides an extension to scenario (D). The definitions of our classifiers do not involve the unknown parameters $a_d$ and $\beta$, in contrast to the ‘ideal’ HC of Donoho & Jin (in press). The proofs are sketched in §7. The detailed proofs can be found in Ingster et al. (2009).

2. Preliminary remarks

In this section we collect some basic remarks on the problem assuming that $f$ is the standard Gaussian density. As a starting point, we discuss some natural limitations for $a_d$.

**Remark 1.** We remark that $a_d$ cannot be too small. Indeed, assume that, instead of the set $U_{\beta,a_d}$, we have only one vector $u = (a_d \varepsilon_1, \ldots, a_d \varepsilon_d)$ with known $\varepsilon_k \in \{0,1\}$. Then we get a familiar problem of classification with two given Gaussian populations. The notion of classification boundary can be defined here...
Sparse high-dimensional vector classes

4433

in the same terms as above, and the explicit form of the boundary can be derived from standard textbook results. It is expressed through the behaviour of \( Q_d^2 \overset{\Delta}{=} a_d^2 \sum_{k=1}^d \varepsilon_k \):

(i) if \( Q_d \to 0 \), then successful classification is impossible, with

\[
\lim \inf_{d \to +\infty} \inf \mathcal{R}(\psi) = \mathcal{R}_{\max};
\]

(ii) if \( Q_d \to +\infty \), then successful classification is realized by the maximum-likelihood classifier \( \psi^* = \mathbb{I}_{\{T^* > 0\}} \), where

\[
T^* = \sum_{k=1, \varepsilon_k=1}^d \left( \frac{Z^k - a_d}{2} \right).
\]

Here and below \( \mathbb{I}_{\{\cdot\}} \) denotes the indicator function.

If we assume that \( \sum_{k=1}^d \varepsilon_k \asymp d^{1-\beta} \), we immediately obtain some consequences for our model defined in §1. We see that successful classification in that model is impossible if \( a_d \) is so small that \( d^{1-\beta} a_d^2 = o(1) \), and it makes sense to consider only such \( a_d \) that

\[
d^{1-\beta} a_d^2 \to +\infty. \tag{2.1}
\]

In particular, for \( \gamma > 1 - \beta \), successful classification is impossible under scenario (C) with \( a_d \asymp \sqrt{(\log d)/m} \).

We shall later see that equation (2.1) is a rough condition, which is necessary but not sufficient for successful classification in the model of §1. For example, in that model with \( \beta \in (0, 1/2] \) and fixed \( m \), the value \( a_d \) should be substantially larger than given by the condition \( d^{1-\beta} a_d^2 \asymp 1 \) (cf. equation (1.5)).

**Remark 2.** Our second remark is that, at the other extreme, for sufficiently large \( a_d \) the problem is trivial. Specifically, non-trivial results can be expected only when the signal-to-noise ratio \( a_d \sqrt{m/\log d} \) is small enough:

\[
x \overset{\Delta}{=} a_d \sqrt{\frac{m}{\log d}} \leq B^*_s,
\]

for a constant \( B^*_s > 0 \). To make this remark clear, assume, for example, that

\[
x > 2\sqrt{2}. \tag{2.3}
\]

Then the problem becomes simple, in the sense that successful classification is easily realizable under (2.3) and the classical condition (2.1). Indeed, take an analogue of the statistic \( T^* \) where \( a_d \) and \( \varepsilon_k \) are replaced by their natural estimators:

\[
T = \sum_{k=1}^d \left( Z^k - SY^k \right) \hat{\varepsilon}_k, \quad \text{with} \quad SY^k = \frac{1}{\sqrt{m}} \sum_{i=1}^m Y^k_i, \quad \hat{\varepsilon}_k = \mathbb{I}_{\{SY^k > \sqrt{2\log d}\}},
\]

and consider the classifier \( \psi = \mathbb{I}_{\{T > 0\}} \). We can write \( SY^k = \varepsilon_k x \sqrt{\log d} + \zeta_k \), where \( \zeta_k \) are independent standard normal random variables. It is well known that \( \max_{k=1, \ldots, d} |\zeta_k| \leq \sqrt{2\log d} \) with probability tending to 1 as \( d \to +\infty \). This and

*Phil. Trans. R. Soc. A* (2009)
equation (2.3) imply that, with probability tending to 1, the vector \((\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_d)\) recovers exactly \((\varepsilon_1, \ldots, \varepsilon_d)\) and the statistic \(T\) coincides with
\[
\hat{T} = \sum_{k=1}^{d} \left( Z_k - \frac{\text{SY}_k}{\sqrt{m}} \right).
\]
Since \(E(u)(\text{SY}_k / \sqrt{m}) = \varepsilon_ka_d\) and \(\text{Var}(u)(\text{SY}_k / \sqrt{m}) = 1/m\), we find that
\[
E_{H_0}(\hat{T}) = -\frac{a_d}{2} \sum_{k=1}^{d} \varepsilon_k \quad \text{and} \quad E_{H_1}(\hat{T}) = \frac{a_d}{2} \sum_{k=1}^{d} \varepsilon_k,
\]
\[
\text{Var}_{H_0}(\hat{T}) = \text{Var}_{H_1}(\hat{T}) = \left(1 + \frac{1}{4m}\right) \sum_{k=1}^{d} \varepsilon_k.
\]
It follows from Chebyshev’s inequality that under condition (2.1) we have
\[
E_{H_0}^{(u)}(\psi) = P_{H_0}^{(u)}(T > 0) \to 0 \quad \text{and} \quad E_{H_1}^{(u)}(1 - \psi) = P_{H_1}^{(u)}(T \leq 0) \to 0, \quad (2.5)
\]
as \(d \to +\infty\). This argument is applicable in the general model of §1 (since the convergence in equation (2.5) is uniform in \(u \in U_{\beta, a_d}\)), implying successful classification by \(\psi\) under conditions (2.1) and (2.3).

**Remark 3.** Let us now discuss a connection between conditions (2.1) and (2.3). First, (2.3) implies (2.1) if \(m\) is not too large, i.e.
\[
m = o(d^{1-\beta} \log d). \quad (2.6)
\]
On the other hand, if \(m\) is very large, i.e.
\[
\exists b > 0 : m \geq bd^{1-\beta} \log d, \quad (2.7)
\]
then we have \(x^2 \geq ba_d^2 d^{1-\beta}\), and condition (2.1) implies (2.3). Thus, the relationship
\[
d^{1-\beta} a_d^2 \asymp 1
\]
determines the classification boundary in the general model of §1, if \(m\) is very large (satisfies (2.7)).

**Remark 4.** Finally, note that conditions (2.3) and (2.2) involve the unknown \(a_d\). In practice, we can replace them by their data-driven counterparts. In fact, set \(M_0 \triangleq \max_{1 \leq k \leq d} \text{SY}_k\). Then with probability tending to 1 as \(d \to \infty\), we have \(M_0 / \sqrt{\log d} \leq x + \sqrt{2}\). Hence, if \(M_0 > 3\sqrt{2 \log d}\), then \(x > 2\sqrt{2}\) with the same probability. It is therefore convenient to consider the following pre-classifier taking values in \(\{0, 1, \text{ND}\}\) (ND means ‘no decision’, i.e. we need to switch to some other classifier):
\[
\psi^{\text{pre}} = \begin{cases} 
0 & \text{if } T \leq 0, \quad M_0 > 3\sqrt{2 \log d}, \\
1 & \text{if } T > 0, \quad M_0 > 3\sqrt{2 \log d}, \\
\text{ND} & \text{if } M_0 \leq 3\sqrt{2 \log d},
\end{cases}
\]
where $T$ is given by (2.4). Suppose that we choose a classifier $\tilde{\psi}$ if $\psi^{\text{pre}}$ takes the value ND. Then the combined classifier is $\tilde{\psi} = \psi^{\text{pre}}I_{\{\psi^{\text{pre}}=\text{ND}\}} + \bar{\psi}I_{\{\psi^{\text{pre}}=\text{ND}\}}$ and

$$E_{H_0}^{(u)}(\tilde{\psi}) \leq E_{H_0}^{(u)}(\tilde{\psi} I_{\{\psi^{\text{pre}}=\text{ND}\}}) + P_{H_0}^{(u)}(\psi^{\text{pre}} = 1) = E_{H_0}^{(u)}(\bar{\psi} I_{\{\psi^{\text{pre}}=\text{ND}\}}) + o(1),$$

$$E_{H_1}^{(u)}(1 - \tilde{\psi}) \leq E_{H_1}^{(u)}((1 - \bar{\psi})I_{\{\psi^{\text{pre}}=\text{ND}\}}) + o(1).$$

Here the $o(1)$ terms are due to the argument in remark 2. We now take $\bar{\psi}$ as one of the classifiers suggested below in this paper. We prove their optimality under assumption (2.2) with any $B_* > 0$. This allows us to treat the terms with $I_{\{\psi^{\text{pre}}=\text{ND}\}}$ in the above relationships. Indeed, with probability tending to 1 as $d \to \infty$, condition $\psi^{\text{pre}} = \text{ND}$ implies (2.2) with $B_* = 3\sqrt{2}$, since $x \leq M_0/\sqrt{\log d}$ with such probability (cf. (7.1) below). In conclusion, $\tilde{\psi}$ has asymptotically the same risk behaviour as $\bar{\psi}$.

The above remarks can be easily extended to the case of Gaussian errors with known variance $\sigma^2 > 0$ by using the normalization $Z_k/\sigma, SY_k/\sigma$. Moreover, they extend to the case of non-Gaussian errors under the Cramér condition and the additional assumption $m/\log d \to +\infty$ (cf. §4d).

### 3. Classifier for moderately sparse vectors

In this section we consider the case of moderately sparse vectors. To simplify the notation, without loss of generality we set $\sigma = 1$. Assume that $R_d = d^{1/2-\beta} a_d$ satisfies

$$\lim_{d \to +\infty} R_d = +\infty \quad (3.1)$$

and consider the classifier based on a linear statistic:

$$\psi^{\text{lin}} = I_{\{T' > 0\}}, \quad T' = \sum_{k=1}^{d} \left( Z^k - \frac{1}{2m} \sum_{i=1}^{m} Y^k_i \right).$$

Note that $T'$ is similar to the statistic $T$ defined in (2.4), the difference being that in $T'$ we do not threshold to estimate the positions of non-zero $\varepsilon_k$. Indeed, here we do not necessarily assume (2.3), and thus there is no guarantee that $\varepsilon_k$ can be correctly recovered.

Assume that $\eta^k_j$ for all $k, j$ are random variables with mean 0 and variance 1. Then the means of $Y^k_i$ and $Z^k$ are $E^{(u)}(Y^k_i) = E_{H_1}^{(u)}(Z^k) = \varepsilon_k a_d$, $E_{H_0}^{(u)}(Z^k) = 0$, their variances are equal to 1, and we have

$$E_{H_0}^{(u)}(T') = -\frac{ad}{2} \sum_{k=1}^{d} \varepsilon_k \quad \text{and} \quad E_{H_1}^{(u)}(T') = \frac{ad}{2} \sum_{k=1}^{d} \varepsilon_k,$$

$$\text{Var}_{H_0}^{(u)}(T') = \text{Var}_{H_1}^{(u)}(T') = d \left( 1 + \frac{1}{4m} \right).$$
We now consider a vector $u \in \mathbb{R}^d$ of the form
\[ u = (u_1, \ldots, u_d): \quad u_k = a_d \varepsilon_k, \quad \varepsilon_k \in \{0, 1\} \quad \text{and} \quad \sum_{k=1}^d \varepsilon_k \geq cd^{1-\beta}. \tag{3.2} \]

By equation (3.2), Chebyshev’s inequality and equation (3.1), we obtain
\[ E_{H_0}^{(u)}(\psi) = P_{H_0}^{(u)}(T' > 0) \leq P_{H_0}^{(u)}(T' - E_{H_0}^{(u)}(T') > cd^{1-\beta} a_d/2) \leq \frac{4d}{(cd^{1-\beta} a_d)^2} \left( 1 + \frac{1}{4m} \right) \to 0, \]
as $d \to +\infty$. An analogous argument yields that $E_{H_1}^{(u)}(1 - \psi) \to 0$. The convergence here is uniform in $u$ satisfying equation (3.2), and thus uniform in $u \in U_{\beta, a_d}$. Therefore, we have the following result.

**Theorem 3.1.** Let $\eta_k^j$ for all $k, j$ be random variables with mean 0 and variance 1. If (3.1) holds, then successful classification is possible and it is realized by the classifier $\psi^{\text{lin}}$.

**Remark.** We have proved theorem 3.1 with the set of vectors $u$ defined by (3.2), which is larger than $U_{\beta, a_d}$. The upper bound on $\sum_k \varepsilon_k$ in the definition of $U_{\beta, a_d}$ is not needed. Also the variances of the $\eta_k^j$ need not be equal to 1. It is easy to see that the result of theorem 3.1 remains valid if these random variables have unknown variances uniformly bounded by an (unknown) constant.

Finally, note that theorem 3.1 is valid for all $\beta \in (0, 1)$. However, for $\beta > 1/2$ its assumption $\lim_{d \to +\infty} R_d = +\infty$ guaranteeing successful classification is much too restrictive when compared with the correct classification boundary that we shall derive in the subsequent sections.

### 4. Classifiers for highly sparse vectors

We now analyse the case of highly sparse vectors, i.e. we suppose that $\beta \in (1/2, 1)$ if $\log m = o(\log d)$, and $\beta^* = \beta/(1 - \gamma) \in (1/2, 1)$ if $\log m \sim \gamma \log d$, $\gamma \in (0, 1)$. The classification boundary in this case is expressed in terms of the function
\[ \phi(\beta) = \begin{cases} \phi_1(\beta) & \text{if } \frac{1}{2} < \beta \leq \frac{3}{4}, \\ \phi_2(\beta) & \text{if } \frac{3}{4} < \beta < 1, \end{cases} \]
where the functions $\phi_1$ and $\phi_2$ are defined in equation (1.9). Note that $\phi_1$ and $\phi_2$ are monotone increasing on $(1/2, 1)$, satisfy $\phi_1(\beta) \leq \phi_2(\beta)$ for all $\beta \in (1/2, 1)$, and the equality $\phi_1(\beta) = \phi_2(\beta)$ ($= 1/\sqrt{2}$) holds if and only if $\beta = 3/4$.

We shall propose some classifiers that attain this boundary.

The following notation will be useful in the sequel:
\[ T_d = \sqrt{\log d}, \quad s = s_d = \frac{a_d}{\sigma T_d}, \tag{4.1} \]
\[ x = s \sqrt{m}, \quad x_0 = \frac{sm}{\sqrt{m + 1}}, \quad x_1 = s \sqrt{m + 1} \quad \text{and} \quad x^* = \frac{x}{1 - \gamma}. \tag{4.2} \]
Clearly, $x_0 < x < x_1$. We allow $s, x, x_0, x_1$ to depend on $d$ but do not indicate this dependence in the notation for the sake of brevity. We shall also suppose throughout that (2.2) holds, so that $x_1 = O(1)$ as $d \to +\infty$.

(a) Classifier for fixed $m$

We now propose optimal classifiers under scenario (A). First, we consider a procedure that attains the classification boundary only for $\beta \in [3/4, 1)$, but has a simple structure. We introduce the statistics

$$M_0 = \max_{1 \leq k \leq d} SY^k \quad \text{and} \quad M = \max_{1 \leq k \leq d} SZ^k,$$

where

$$SY^k = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} Y_i^k \quad \text{and} \quad SZ^k = \frac{1}{\sqrt{m} + 1} \left( Z^k + \sum_{i=1}^{m} Y_i^k \right).$$

Define

$$A_M = \frac{M}{\max(\sqrt{2} \sigma T_d, M_0)}.$$

Taking a small $c_0 > 0$, consider the classifier of the form,

$$\psi^{\max} = I_{\{A_M > 1 + c_0\}}.$$

**Theorem 4.1.** Consider scenario (A). Let $\beta \in (0, 1)$ and (2.2) hold. Then, for any $c_0 > 0$,

$$\lim_{d \to +\infty} \sup_{u \in U_{\beta, ad}} E_{H_0}^{(u)}(\psi^{\max}) = 0. \quad (4.4)$$

If $\limsup_{d \to +\infty} x_1 < \phi_2(\beta)$, then, for any $c_0 > 0$,

$$\lim_{d \to +\infty} \sup_{u \in U_{\beta, ad}} E_{H_1}^{(u)}(\psi^{\max}) = 0. \quad (4.5)$$

If $\liminf_{d \to +\infty} x_1 > \phi_2(\beta)$, then there exists $c_0 > 0$ such that

$$\lim_{d \to +\infty} \sup_{u \in U_{\beta, ad}} E_{H_1}^{(u)}(1 - \psi^{\max}) = 0. \quad (4.6)$$

Theorem 4.1 (cf. equations (4.4) and (4.6) and the fact that $\phi(\beta) = \phi_2(\beta)$ for $\beta \in [3/4, 1]$) and theorem 6.3 below imply that $\psi^{\max}$ attains the classification boundary for $\beta \in [3/4, 1)$. On the other hand, equation (4.5) implies that, for $\beta \in (1/2, 3/4)$ (where $\phi(\beta) = \phi_1(\beta) < \phi_2(\beta)$), the classifier $\psi^{\max}$ does not do the correct job. Its maximal risk $R_M$ is asymptotically 1, which is larger than the risk 1/2 of the simple random guess. We therefore introduce another classifier.
that has, however, a more involved structure. Consider the statistics

\[ L_0(t) = \sum_{k=1}^{d} (\mathbb{I}_{[SY^k > \sigma T_d]} - \Phi(-tT_d)), \quad \Delta_0(t) = \frac{L_0(t)}{\sqrt{d\Phi(-tT_d)}}, \]

\[ L(t) = \sum_{k=1}^{d} (\mathbb{I}_{[SZ^k > \sigma T_d]} - \Phi(-tT_d)), \quad \Delta(t) = \frac{L(t)}{\sqrt{d\Phi(-tT_d)}}, \]

where \( t \in \mathbb{R} \), \( \Phi \) is the standard normal cumulative distribution function and the statistics \( SY^k \), \( SZ^k \) are defined in equation (4.3). Consider the grid

\[ t_l = lh, \quad l = 1, \ldots, N \quad \text{and} \quad t_N = \sqrt{2} \sigma, \]

with a step \( h > 0 \) depending on \( d \) and such that \( h = o(1), \quad T_d h \to +\infty \). This implies that \( 1 \ll N \ll T_d \) as \( d \to +\infty \) (here and below \( v_d \ll w_d \) for \( v_d > 0 \) and \( w_d > 0 \) depending on \( d \) means that \( \lim_{d \to +\infty} v_d / w_d = 0 \)). Set

\[ \Delta_0 = \max_{1 \leq l \leq N} \Delta_0(t_l), \quad \Delta = \max_{1 \leq l \leq N} \Delta(t_l) \quad \text{and} \quad \Lambda^* = \frac{\Delta}{H + \Delta_0}, \]

where \( H = H_d \) is such that

\[ d^{bh} \ll H \ll d^B, \]

for any \( B > 0, \quad b > 0 \) and any \( d > d_0(B, b) \), where \( d_0(B, b) \) is a constant depending only on \( B \) and \( b \) (such an \( H \) can always be determined depending on the choice of \( h \)). Now consider the classifier of the form

\[ \psi^*_m = \mathbb{I}_{\{\Lambda^* > H\}}. \]

**Theorem 4.2.** Consider scenario (A) with \( \beta \in (1/2, 1) \) and assume (2.2). Then

\[ \lim_{d \to +\infty} \sup_{u \in U_{\beta, ad}} E_{H_0}^{(u)}(\psi^*_m) = 0. \]

If \( \liminf_{d \to +\infty} x_1 > \phi(\beta) \) and \( \limsup_{d \to +\infty} x_0 < \sqrt{2} \), then

\[ \lim_{d \to +\infty} \sup_{u \in U_{\beta, ad}} E_{H_1}^{(u)}(1 - \psi^*_m) = 0. \]

Theorems 4.1, 4.2 and 6.3 below show that the classification boundary for highly sparse vectors (i.e. for \( \beta \in (1/2, 1) \)) is given by (1.7). Furthermore, the classifier \( \psi^*_m \) is optimal (attains the classification boundary) for \( \beta \in (1/2, 1) \), except for the case \( \limsup_{d \to +\infty} x_0 \geq \sqrt{2} \), which is already covered by the classifier \( \psi^{max} \). Indeed, \( x_0 \geq \sqrt{2} \) implies that \( x_1 \geq \sqrt{2}(1 + 1/m) > \phi_2(\beta) \) for all \( \beta \in (1/2, 1) \).

(b) **Classifier for** \( m \to +\infty, \quad \log m = o(\log d) \)

In this subsection, we analyse scenario (B). Then \( m = m_d \to +\infty, \quad \log m = o(\log d) \) as \( d \to +\infty \) and the classifier \( \psi^*_m \) is not, in general, optimal. Nevertheless, we propose another classifier \( \psi^*_\infty \) that essentially attains the same classification.

*Phil. Trans. R. Soc. A* (2009)
boundary as in §4a above. We introduce the statistics

$$\Delta(t) = \frac{1}{\sigma \sqrt{d} \Phi(-t \sqrt{d})} \sum_{k=1}^{d} Z_k \mathbb{I}_{[S_Y^k > \sigma t T_d]}$$

and

$$\Delta = \max_{1 \leq l \leq N} \Delta(t_l),$$

where the maximum is taken over the grid (4.7). Here and below we use the same notation $\Delta(t)$, $\Delta$ as previously for different ratio statistics, since it causes no ambiguity. Set also

$$\Delta_* = \sum_{k=1}^{d} \mathbb{I}_{[S_Y^k > \sqrt{2} \sigma T_d]},$$

and define

$$\Lambda_*^\alpha = \frac{\Delta}{\sqrt{H + \Delta_*}}$$

and

$$\psi_*^\alpha = \mathbb{I}_{[\Lambda_*^\alpha > H]},$$

(4.11)

where $H$ satisfies (4.8).

**Theorem 4.3.** Consider scenario (B). Let $\beta \in (1/2, 1)$ and let (2.2) hold. Then

$$\lim_{d \to +\infty} \sup_{u \in U_{\beta, a d}} E_{H_0}^{(u)}(\psi_*^\alpha) = 0. \quad (4.12)$$

If $\liminf_{d \to +\infty} x_* > \phi(\beta^*)$, then

$$\lim_{d \to +\infty} \sup_{u \in U_{\beta, a d}} E_{H_1}^{(u)}(1 - \psi_*^\alpha) = 0. \quad (4.13)$$

(c) **Classifier for scenario (C)**

We now suggest an asymptotically optimal classifier for scenario (C). For $t \geq 0$, we introduce the statistics

$$L^1(t) = \sum_{k=1}^{d} Z_k \mathbb{I}_{[S_Y^k > \sigma t T_d]}; \quad L^0(t) = \sum_{k=1}^{d} \mathbb{I}_{[S_Y^k > \sigma t T_d]} \quad \text{and} \quad \Delta(t) = \frac{L^1(t)}{\sigma \sqrt{N^2 + L^0(t)}}.$$

Take a grid $t_1, \ldots, t_N$ of the form (4.7) and define the classifier

$$\psi_* = \mathbb{I}_{[\Lambda_* > 4 N]}, \quad \text{where} \quad \Delta = \max_{1 \leq l \leq N} \Delta(t_l).$$

**Theorem 4.4.** Consider scenario (C) with $a_d = \sigma x \sqrt{\log d}/m$. Let $\beta^* = \beta/(1 - \gamma) \in (1/2, 1)$ and let (2.2) hold. Then

$$\lim_{d \to +\infty} \sup_{u \in U_{\beta, a d}} E_{H_0}^{(u)}(\psi_*^\alpha) = 0. \quad (4.14)$$

If $\liminf_{d \to +\infty} x_* > \phi(\beta^*)$, then

$$\lim_{d \to +\infty} \sup_{u \in U_{\beta, a d}} E_{H_1}^{(u)}(1 - \psi_*^\alpha) = 0. \quad (4.15)$$
Extensions for non-Gaussian noise

We now discuss an extension of our results to scenario (D). The remarks on
the pre-classifier $\psi^{\text{pre}}$ in §2 and the proofs of the upper bounds are based only
on the constraint (2.2) and the following property of the tails of the Gaussian
distribution:

$$\log P \left( S \zeta^m > \sigma t \right) \sim -\frac{t^2}{2}, \quad t \in [U_0, U_1] \quad \text{for} \quad U_0 \to +\infty \quad \text{and} \quad U_1 = O(T_d). \quad (4.16)$$

Here $S \zeta^m = (1/\sqrt{m}) \sum_{i=1}^m \zeta_i$, and $\zeta_i$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables.

Indeed, in the proof of theorem 4.4 (see §7 below), we can write $\Phi(-t T_d)$ as
$$P \left( S \zeta^m > \sigma t T_d \right) = A_d d^{-(t_+)^2/2}, \quad t_+ = \max(0, t),$$
where $A_d$ satisfies, for $t_+ = O(1)$,
$$|\log A_d| = o(\log d), \quad (4.17)$$
as $d \to +\infty$. This is exactly the relationship
$$\Phi(-t T_d) = A_d d^{-(t_+)^2/2}, \quad t_+ = \max(t, 0), \quad (4.18)$$
which is also the only property of the noise distribution needed for the proof of theorem 4.4.

If $m$ is large enough, relationship (4.16) holds not only for the Gaussian $\zeta_i$. It
suffices to have the i.i.d. $\zeta_i$ with $E\zeta_i = 0$, $E(\zeta_i^2) = \sigma^2 > 0$ satisfying the Cramér
condition:
$$\exists h_0 > 0: \quad E(e^{h\zeta_i}) < +\infty, \quad \forall h \in (-h_0, h_0),$$
and $m \gg \log d$. In fact, using theorem 5.23 in Petrov (1995) we get that, under
the Cramér condition and for $t = o(\sqrt{m})$,
$$P \left( S \zeta^m > \sigma t \right) = \Phi(-t) \exp \left( \frac{t^3}{\sqrt{m}} \lambda \left( \frac{t}{\sqrt{m}} \right) \right) \left\{ 1 + O \left( \frac{t + 1}{m} \right) \right\},$$
where $\lambda(t)$ is the Cramér series. Inserting here the expression for the Cramér series
and the relationship $\log \Phi(-t) = -t^2/2 - \log t + O(1)$ as $t \to +\infty$, we obtain
$$\log P \left( S \zeta^m > \sigma t \right) = -\frac{t^2}{2} \left( 1 + O \left( \frac{t}{\sqrt{m}} \right) + o(1) \right) \sim -\frac{t^2}{2},$$
as $t \to +\infty$, $t = o(\sqrt{m})$. These remarks allow us to follow the proof of theorem 4.4
leading to the next result.

**Theorem 4.5.** Consider scenario (D) with $a_d = \sigma x \sqrt{(\log d)/m}$. Let $\beta^* = \beta/(1 - \gamma) \in (1/2, 1)$ and let (2.2) hold. Then
$$\lim_{d \to +\infty} \sup_{u \in \mathcal{U}_{\beta,a_d}} E_{h_0}^{(u)}(\psi^\infty) = 0. \quad (4.19)$$
If \( \liminf_{d \to +\infty} x^* > \phi(\beta^*) \), then

\[
\lim_{d \to +\infty} \sup_{u \in U_{\beta, a_d}} E_{H_1}^{(u)} (1 - \psi_{\infty}) = 0.
\]

(4.20)

5. Adaptive procedures

We have proposed several classifiers in \( \S \S 3 \) and 4. We would like to stress that their definitions do not involve the unknown parameters \( a_d \) and \( \beta \), in contrast to the ‘ideal’ classifiers of Donoho & Jin (in press). We prove the optimality of our classifiers (i.e. the fact that they attain the boundary) under different conditions on \( a_d \) and \( \beta \). This can be easily generalized to obtain an adaptive procedure, i.e. the one attaining the classification boundary simultaneously for all the considered domains of \( a_d \) and \( \beta \). It suffices to combine the classifiers as follows. We start with the pre-classifier \( \psi_{\text{pre}} \). If it outputs ‘no decision’ (ND) and \( m \) is not too large (say, \( m \leq 10 \rightarrow 30 \)), then we combine the classifiers \( \psi_{\text{lin}}, \psi_{\text{max}} \) and \( \psi_{m} \) using the Bonferroni device, i.e. our final classifier is \( \max(\psi_{\text{lin}}, \psi_{\text{max}}, \psi_{m}) \). In other words, we allocate \( Z \) to the \( P_Y \) population if and only if it is allocated to \( P_Y \) by at least one of the three classifiers. Analogously, if \( m \to +\infty \) (in practice, \( m > 30 \)), then for \( \psi_{\text{pre}} = \text{ND} \) it suffices to classify by \( \max(\psi_{\text{lin}}, \psi_{\infty}) \) or by \( \max(\psi_{\text{lin}}, \psi_{\infty}) \).

6. Lower bounds

We now prove the lower bounds on the risks to show that relationships (1.5) and (1.10) determine the classification boundaries for moderately and highly sparse vectors, respectively.

(a) Lower bound for moderately sparse vectors

The lower bound for moderately sparse vectors is based on the next theorem. For \( a > 0 \), \( t \in \mathbb{R} \), set

\[
\ell_a(t) = \frac{f(t - a)}{f(t)} \quad \text{and} \quad D_a = \int \ell_a^2(t) f(t) \, dt,
\]

and

\[
D_d(m, a, \beta) = d^{1-2\beta} D_a^{m} (D_a - 1).
\]

**Theorem 6.1.** Let either \( m \geq 1 \) be fixed or \( m = m_d \to +\infty \). If

\[
\lim_{d \to +\infty} D_d(m, a_d, \beta) = 0,
\]

then successful classification is impossible.

**Corollary 6.2.** Let \( f \) be the density of the standard normal distribution. If

\[
\lim_{d \to +\infty} R_d = 0,
\]

then successful classification is impossible for \( \beta \in (0, 1/2] \) and \( m \) fixed or for \( \beta \in (0, 1/2) \) and \( m = m_d \to +\infty \) such that \( m = O(d^{1-2\beta}) \).
Other choices of for non-Gaussian noise with finite Fisher information. Indeed, assume that is the Fisher information of with no exception, so that there is no classification boundary in this range of . Thus, if \( \gamma = \text{constant} \) for instance, in Ibragimov & Has’minski (1981)). Under assumption (6.4), we have
\[
\gamma = \frac{\beta}{2} \leq \frac{1}{2},
\]
which means that the Fisher information is large, i.e. \( \log \gamma \sim m \). For example, consider \( \gamma = \frac{1}{2} \) and \( m = m_d \rightarrow +\infty \) such that \( m = O(d^{1-2\beta}) \), then \( ma^2 = O(R_d) = o(1) \).

Note that, for \( m \sim d^\beta \) with \( 0 < \gamma < 1 \), the condition \( m = O(d^{1-2\beta}) \) in theorem 6.1 becomes \( \beta^* \triangleq \beta/(1 - \gamma) \leq 1/2 \).

Remark. Relationship (6.3) is valid for a larger class of noise distributions, e.g. for non-Gaussian noise with finite Fisher information. Indeed, assume that \( \ell_a(t) \) is \( L_2(f) \)-differentiable at point \( a = 0 \), i.e. there exists a function \( \ell'(\cdot) \) such that
\[
\| \ell_a(\cdot) - 1 - a\ell'(\cdot) \|_f = o(a) \quad \text{and} \quad 0 < \| \ell'(\cdot) \|_f < +\infty,
\]
where \( \| g(\cdot) \|_f^2 = \int_R g^2(x) f(x) \, dx \). Observe that
\[
\| \ell'(\cdot) \|_f^2 = \int_R \frac{(f'(x))^2}{f(x)} \, dx \triangleq I(f)
\]
is the Fisher information of \( f \) (with \( f' \) defined in a somewhat stronger sense than, for instance, in Ibragimov & Has’minski (1981)). Under assumption (6.4), we have
\[
D_a = 1 + \| \ell_a(\cdot) - 1 \|_f^2, \quad \| \ell_a(\cdot) - 1 \|_f^2 = a^2 (I(f) + o(1)) \quad \text{as} \quad a \rightarrow 0.
\]
Combining the remarks mentioned earlier with theorems 3.1 and 6.1, we see that relationship (1.5) determines the classification boundary for \( \beta^* \in (0, 1/2] \) when \( m \) is fixed or \( m \rightarrow +\infty \), \( m \sim d^\gamma \), \( \gamma \in [0, 1] \), if the errors have mean 0, finite variance and finite Fisher information.

As corollaries of theorems 3.1 and 6.1, we can establish classification boundaries for particular choices of \( a_d \). Recall that non-trivial results can be expected only if \( a_d \) satisfies (2.2). For instance, consider \( a_d = d^{-s} \) with some \( s > 0 \). Then, for fixed \( m \), the classification boundary in the region \( \beta \in (0, 1/2] \) is given by \( s = \beta - 1/2 \), i.e. successful classification is possible if \( s < 1/2 - \beta \), and is impossible if \( s > 1/2 - \beta \). Other choices of \( a_d \) appear to be less interesting when \( \beta \in (0, 1/2] \). For example, in the next section we consider the sequence \( a_d = s\sigma \sqrt{(\log d)/m} \) with some \( s > 0 \). If \( a_d \) is chosen in this way, successful classification is possible for all \( \beta \in (0, 1/2] \) with no exception, so that there is no classification boundary in this range of \( \beta \).

(b) Lower bound for highly sparse vectors

**Theorem 6.3.** Let the noise density \( f \) be Gaussian \( N(0, \sigma^2) \), \( \sigma^2 > 0 \). Assume that \( \beta \in (1/2, 1) \) and \( \limsup_{d \rightarrow +\infty} x_1 < \phi(\beta) \). Then successful classification is impossible for fixed \( m \) and for \( m = m_d \rightarrow +\infty \).

Though theorem 6.3 is valid with no restriction on \( m \), it does not provide a correct classification boundary if \( m \) is large, i.e. \( \log m \sim \gamma \log d \), \( \gamma \in (0, 1) \), as in scenarios (C) and (D). The correct lower bound for large \( m \) is given in the next theorem.
**Theorem 6.4.** Consider scenario (C) with $\beta^* = \beta / (1 - \gamma) \in (1/2, 1)$ and 

$$a_d = \sigma x \sqrt{\frac{(\log d)}{m}}.$$ 

Assume that $\limsup_{d \to +\infty} x^* < \phi(\beta^*)$. Then successful classification is impossible.

Recall that, by an elementary argument, under scenario (C) and for $a_d$ as in theorem 6.4, successful classification is impossible if $\beta > 1 - \gamma$ (cf. remark after (2.1)). This is the reason why in theorem 6.4 we consider only $\beta < 1 - \gamma$.

### 7. Sketch of the proofs

#### (a) Proofs of upper bounds

Theorem 4.1 follows from the following relationships: $\forall \delta > 0$ uniformly in $u \in U_{\beta, a_d}$,

$$P^{(u)}(|M_0 - h(x)\sqrt{\log d}| > \delta) \to 0$$

and

$$P^{(u)}_{H_s}(|M - h(x_s)\sqrt{\log d}| > \delta) \to 0, \quad s = 0, 1, \quad (7.1)$$

where $h(t) = \max(\sqrt{2}, t + \sqrt{2(1 - \beta)})$.

The proofs of theorems 4.2–4.4 are based on evaluation of the expectations and variances of the statistics under consideration. For theorem 4.2, we get for $x, x_s > b$ and all $d$ large enough uniformly in $u \in U_{\beta, a_d}$ with log-factor $A_d$,

$$E^{(u)}\Delta_0(t) = A_d d^{1/2 - \beta} d^{t^2/4 - (t - x)^2/2}, \quad E^{(u)}_{H_s}\Delta(t) = A_d d^{1/2 - \beta} d^{t^2/4 - (t - x_s)^2/2},$$

where $s = 0, 1$. Since the maximum of $t^2/4 - ((t - x_x)^2)/2$ in $0 \leq t \leq \sqrt{2}$ is attained either at $t = 2x$ when $0 < x \leq 1/\sqrt{2}$, or at $t = \sqrt{2}$ when $x > 1/\sqrt{2}$, we have

$$\max_{h \leq t \leq \sqrt{2}} E^{(u)}\Delta_0(t) = \begin{cases} 
A_d d^{1/2 - \beta + x^2/2}, & x \leq 1/\sqrt{2}, \\
A_d d^{1/2 - \beta - ((\sqrt{2} - x)^2)/2}, & x > 1/\sqrt{2}. 
\end{cases} \quad (7.2)$$

For the variances, we have

$$\text{Var}^{(u)}(\Delta_0(t)) \leq A_d (1 + d^{-\beta + (t^2 - (t - x)^2)/2}) \leq A_d (1 + d^{1-\beta - (\sqrt{2} - x)^2/2})$$

$$\leq \begin{cases} 
A_d, & x \leq \phi_2(\beta), \\
A_d d^{1-\beta - (\sqrt{2} - x)^2/2}, & x > \phi_2(\beta), 
\end{cases}$$

$$\text{Var}^{(u)}_{H_s}(\Delta(t)) \leq A_d (1 + d^{-\beta + (t^2 - (t - x_s)^2)/2}) \leq A_d (1 + d^{1-\beta - (\sqrt{2} - x_s)^2/2})$$

$$\leq \begin{cases} 
A_d, & x_s \leq \phi_2(\beta), \\
A_d d^{1-\beta - (\sqrt{2} - x_s)^2/2}, & x_s > \phi_2(\beta), 
\end{cases}$$

where $s = 0, 1$. Take $N_0 > 0$ such that $N_0^2 \asymp T_d \gg N$. By Chebyshev’s inequality, for each $l = 1, \ldots, N$, each $u \in U_{\beta, a_d}$ and $s = 0, 1$, with $P^{(u)}_{H_s}$ probability greater.
than \(1 - 1/N_0^2\), we have, for \(D_s^2 = \max_{0 \leq t \leq \sqrt{2}} \text{Var}^{(u)}_{H_s}\Delta_0(t)\),
\[
E^{(u)}_{H_s}\Delta_0(t) - N_0D_s \leq \Delta_0(t) \leq E^{(u)}_{H_s}\Delta_0(t) + N_0D_s,
\]
and these inequalities are also valid for \(\Delta(\cdot)\) instead of \(\Delta_0(\cdot)\). All these inequalities (with \(\Delta(\cdot)\) and \(\Delta_0(\cdot)\)) simultaneously hold with probability greater than \(1 - 2N_0^{-2}N \to 1\) (uniformly in \(u \in U_{\beta,a_d}\)). On this event of high probability we can evaluate \(\Delta_0\) and \(\Delta\) by taking the maxima of the expectations and comparing them with the maxima of the square roots of the variances. Proceeding in this way and using the bounds obtained above, we find
\[
\Delta_0 = \begin{cases} 
O(A_d), & x \leq \phi_1(\beta), \beta \leq 3/4 \lor x < \phi_2(\beta), \beta > 3/4, \\
d^{1/2 - 2/\beta + x^2/2 + O(h)}, & 1/\sqrt{2} \leq x \geq \phi_1(\beta), \beta \leq 3/4, \\
d^{1 - \beta - ((\sqrt{2} - x)^2/2 + O(h))}, & x > 1/\sqrt{2}, \beta \leq 3/4 \lor x \geq \phi_2(\beta), \beta > 3/4,
\end{cases}
\]
with \(P^{(u)}\) probability tending to 1 as \(d \to +\infty\), and, for \(s = 0,1\),
\[
\Delta = \begin{cases} 
O(A_d), & x_s \leq \phi_1(\beta), \beta \leq 3/4 \lor x_s < \phi_2(\beta), \beta > 3/4, \\
d^{1/2 - 2/\beta + x_s^2/2 + O(h)}, & 1/\sqrt{2} \leq x_s \geq \phi_1(\beta), \beta \leq 3/4, \\
d^{1 - \beta - ((\sqrt{2} - x_s)^2/2 + O(h))}, & x_s > 1/\sqrt{2}, \beta \leq 3/4 \lor x_s \geq \phi_2(\beta), \beta > 3/4,
\end{cases}
\]
with \(P^{(u)}_{H_s}\) probability tending to 1 as \(d \to +\infty\) (the convergence of all the probabilities is uniform in \(u \in U_{\beta,a_d}\)). Using these relationships we get the results of theorem 4.2.

For theorem 4.3, we have \(E^{(u)}_{H_0}\Delta(t) = 0\), \(\text{Var}^{(u)}_{H_0}\Delta(t) = 1 + A_d d^{1/2 - 2/\beta + t^2/2 - ((t - x)^2/2),}
\]
\[
D_0^2(x, \beta) = \max_{0 \leq t \leq \sqrt{2}} \text{Var}^{(u)}_{H_0}\Delta(t) = 1 + A_d \begin{cases} 
O(1), & x \leq \phi_2(\beta), \\
d^{1 - \beta - ((\sqrt{2} - x)^2/2)}, & x \geq \phi_2(\beta).
\end{cases}
\]
Next, \(E^{(u)}_{H_1}\Delta(t) = a_d A_d d^{1/2 - 2/\beta + t^2/2 - ((t - x)^2/2),}\) for \(E(x, \beta) = \max_{0 \leq t \leq \sqrt{2}} E^{(u)}_{H_1}\Delta(t)\) we have
\[
E(x, \beta) = a_d A_d \begin{cases} 
O(1), & x \leq \phi_1(\beta), \beta \leq 3/4 \lor x \leq \phi_2(\beta), \beta \geq 3/4, \\
d^{1/2 - 2/\beta + x^2/2}, & 1/\sqrt{2} \leq x \geq \phi_1(\beta), \beta \leq 3/4, \\
d^{1 - \beta - ((\sqrt{2} - x)^2/2)}, & x > \phi_2(\beta), \beta \geq 3/4 \lor x \geq 1/\sqrt{2}, \beta \leq 3/4,
\end{cases}
\]
Analogously for \(D_1^2(x, \beta) = \max_{0 \leq t \leq \sqrt{2}} \text{Var}^{(u)}_{H_1}\Delta(t)\) we have
\[
D_1^2(x, \beta) = 1 + A_d (1 + a_d^2) \begin{cases} 
O(1), & x \leq \phi_2(\beta), \\
d^{1 - \beta - ((\sqrt{2} - x)^2/2)}, & x \geq \phi_2(\beta).
\end{cases}
\]
Using these relationships, similar to the above, we obtain theorem 4.3.

For theorem 4.4, we have
\[
E^{(u)}_{H_0} L^1(t) = 0, \quad E(t) = E^{(u)}_{H_1} L^1(t) = a_d d^{1 - \beta} \Phi(a_d \sqrt{m} - tT_d) = A_a d^{1 - \beta - (t - x)^2/2},
\]
and, for \(R(t) = \max(d\Phi(-tT_d), dp\Phi(a_d \sqrt{m} - tT_d))\), we have
\[
\text{Var}^{(u)}_{H_0} L^1(t) \leq 2R(t), \quad \text{Var}^{(u)}_{H_1} L^1(t) \leq 3R(t), \quad \text{Var}^{(u)} L^0(t) \leq E^{(u)} L^0(t) \leq R(t).
\]
Thus, for all \( l = 1, \ldots, N \), with \( P_{H_0}^{(u)} \) probability tending to 1, the statistics \( L^1(t) \) belong to the intervals \([-N \sqrt{2R(t)}, N \sqrt{2R(t)}]\) and, with \( P_{H_1}^{(u)} \) probability tending to 1, they belong to the intervals \([E(t) - N \sqrt{3R(t)}, E(t) + N \sqrt{3R(t)}]\), and the statistics \( E^0(t) \) belong to \([0, R(t) + N \sqrt{R(t)}]\) with probability tending to 1.

Consider the ratios \( \Delta(t) \), \( l = 1, \ldots, N \). We check that, for all \( l = 1, \ldots, N \) such that \( R(t) > 4N^2 \), with \( P_{H_0}^{(u)} \) probability tending to 1, \( \Delta(t) < 4N \) for \( s = 0 \), and \( \Delta(t) \geq E(t)/(2 \sqrt{R(t)}) - N \) for \( s = 1 \). Thus, \( E_{H_0}^{(s)} \psi = P_{H_0}^{(s)}(\Delta > 4N) \rightarrow 0 \), uniformly over \( u \in U_{\beta, ad} \), and it suffices to show that \( \max_t E(t)/(2 \sqrt{R(t)}) > d^* \), for some \( \tau > 0 \). Let us study the ratio \( E(t)/\sqrt{R(t)} \) assuming that \( x^* > \phi(\beta^*) \). We have, with a logarithmic factor \( A_d \),

\[
E(t)/\sqrt{R(t)} = A_d d^{s(t)}, \quad s(t) = -\varphi/2 + 1/2 - \beta + 1/2 \min \left( t^2/2 - (t - x)^2, \frac{\beta - (t - x)^2}{2} \right).
\]

Set \( t_0 = x/2 + \beta/x \). We check that

\[
s^* = \max_{0 \leq t \leq \sqrt{2}} s(t) \geq -\varphi/2 + 1/2 + \left\{ \begin{array}{ll}
-\beta/2, & \text{if } x \geq t_0, \\
-t_0^2/4, & \text{if } x \leq t_0 \leq 2x, \\
-\beta + x^2/2, & \text{if } 2x \leq t_0.
\end{array} \right.
\]

These relationships imply that \( s^* > \tau \) for some \( \tau > 0 \) as \( x^* > \phi(\beta^*) \). Since \( s(\cdot) \) is a Lipschitz function, we can replace the maximum over the interval \([0, \sqrt{2}]\) by the maximum over our grid with step \( h \), inducing the error of order \( O(h) \). This yields theorem 4.4.

(b) Proofs of lower bounds

We bound the maximal risk from below by the Bayesian risk corresponding to the mixtures \( P_{H_1} \) of the measures \( P_{H_1}^{(l)}, l = 0, 1 \), over the product prior measure \( \mu^d(du) \), where \( \mu = (1 - p)\delta_0 + p\delta_{ad} \), \( \delta_a \) is the Dirac mass at point \( a \) and \( p = d^{-\beta} \). We show that \( \mu^d(U_{\beta, ad}) \rightarrow 1 \). Then it suffices to verify that

\[
E_{H_0} \left| \frac{dP_{H_1}}{dP_{H_0}} - 1 \right| \rightarrow 0. \tag{7.4}
\]

Set \( L(Y, Z) = dP_{H_1}/P_{H_0}, \ell_a(t) = f(t - a)/f(t) \). We have \( L(Y, Z) = \prod_{k=1}^d (1 + \Delta_k) \) where

\[
\Delta_k = \frac{pL(Y^k)(l_{ad}(Z^k) - 1)}{1 + p(L(Y^k) - 1)} \quad \text{and} \quad L(Y^k) = \prod_{j=1}^m l_{ad}(Y_j^k).
\]

Theorem 6.1 follows from the relationships

\[
\left( E_{H_0} \left| \frac{dP_{H_1}}{P_{H_0}} - 1 \right| \right)^2 \leq E_{H_0}(L(Y, Z) - 1)^2 = 1 + E_{H_0}L^2(Y, Z)
\]

Phil. Trans. R. Soc. A (2009)
and (with the measure $P_{H_{0,k}}$ corresponding to the $k$th coordinate)

$$
E_{H_0}(L^2(Y, Z)) = \prod_{k=1}^{d} (1 + E_{H_0,k} \Delta_k^2) \leq \exp \left( \sum_{k=1}^{d} E_{H_0,k} \Delta_k^2 \right)
$$

$$
= \exp \left( \sum_{k=1}^{d} E_{0,k} \left( \Delta_k^2 \frac{dP_{H_0,k}}{dP_{0,k}} \right) \right)
$$

$$
\leq \exp \left( \frac{p^2}{1-p} \sum_{k=1}^{d} E_{0,k} L^2(Y_k) E_{0,k}(\ell_{ad}(Z^k) - 1)^2 \right)
$$

$$
= \exp \left( \frac{dp^2 D_{ad}^m(D_{ad} - 1)}{1-p} \right). \tag{7.5}
$$

For the proof of theorem 6.3, without loss of generality consider $\sigma = 1$ and introduce the events

$$
A_{SY,k} = \left\{ \sum_{j=1}^{m} Y_j^k < \sqrt{2m \log d} \right\} \quad \text{and} \quad A_{SZ,k} = \left\{ Z^k + \sum_{j=1}^{m} Y_j^k < \sqrt{2(m+1) \log d} \right\},
$$

and $A_{SY} = \bigcap_{k=1}^{d} A_{SY,k}$ and $A_{SZ} = \bigcap_{k=1}^{d} A_{SZ,k}$. We show that

$$
P_{H_s}(A_{SY}) \to 1 \quad \text{and} \quad P_{H_s}(A_{SZ}) \to 1, \quad s = 0, 1. \tag{7.6}
$$

We repeat calculations similar to (7.5) for the truncated version of the statistics

$$
\hat{L}(Y, Z) = L(Y, Z) 1_{A_{SY} \cap A_{SZ}}. \quad \text{The relationships (7.6) then yield} \quad E_{H_0}\hat{L}(Y, Z) \to 1
$$

and we get

$$
E_{H_0}(\hat{L}^2(Y, Z)) \leq \exp \left( \frac{dp^2 A}{1-p} + o(1) \right),
$$

where

$$
A = E_{0,1}[L^2(Y^1)(\ell_{ad}(Z^1) - 1)^21_{A_{SY,1} \cap A_{SZ,1}}] \leq B + C,
$$

$$
B = e^{(m+1)a^2} \Phi(\sqrt{2 \log d - 2a_d \sqrt{m + 1}}), \quad C = e^{ma^2} \Phi(\sqrt{2 \log d - 2a_d \sqrt{m}}) \leq B.
$$

Finally we observe, for $1/\sqrt{2} < x_1 \leq \sqrt{2}$, that

$$
dp^2 B \propto (\log d)^{-1/2} d^{-2\beta + 2\sqrt{2}x_1 - x_1^2} = (\log d)^{-1/2} d^{2(1-\beta) - (\sqrt{2} - x_1)^2},
$$

and the exponent is negative under the assumption of theorem 6.3.

For theorem 6.4 in view of the first two lines of (7.5), it suffices to show that

$$
\sum_{k=1}^{d} E_{H_0,k} \Delta_k^2 = dE_{H_0,1} \Delta_1^2 = o(1). \quad \text{Set} \quad \Delta_Y = pL(Y^1)/1-p + pL(Y^1) \quad \text{and observe} \quad \Delta_Y \leq (1-p)^{-1} \min(1, pL(Y^1)).
$$

Next, we have

$$
L(Y^1) = \exp(-ma^2/2 + \sqrt{maSY^1}) = d^{-x^2/2 + xSY^1/\sqrt{\log d}}, \quad SY^1 = m^{-1/2} \sum_{i=1}^{m} Y^1_i.
$$

---

*Phil. Trans. R. Soc. A (2009)*
Take a threshold $H_s = t \sqrt{\log d}$ such that $pL(H_s) = 1$, i.e. $t = x/2 + \beta/x$. Then $pL(Y^1) < 1$ (respectively, $pL(Y^1) > 1$) is equivalent to $SY^1 < H_s$ (respectively, $SY^1 > H_s$). We find $E_{H_{0,1}}(\ell_a(Z^1) - 1)^2 = e_{0,a}^2 - 1 \leq c_0 a_d^2$, and
\[
E_{H_{0,1}} \Delta^2_1 = E_{H_{0,1}}(\ell_a(Z^1) - 1)^2 E_P \Delta^2_{Y^1} \leq c_0 a_d^2 E_P \Delta^2_{Y^1} \\
\leq \frac{c_0 a_d^2}{1 - p} (p^2 E_P (L^2(Y^1) I_{pL(Y^1) \leq 1}) + P(pL(Y^1) > 1)),
\]
where $P = (1 - p)P_0 + pP_d$, $a = a_d$ for brevity, and $P_d$ is the Gaussian measure with the density $f_d(\cdot)$. Evaluating these items, we get, with log-factor $A_d$, 
\[
dE_{H_{0,1}} \Delta^2_1 = A_d d_v^d \quad \text{and} \quad v_d = -\gamma + 1 + \begin{cases} -\beta & x^2 \geq 2\beta, \\
-\beta & 2\beta/3 \leq x^2 \leq 2\beta, \\
-2\beta + x^2 & 0 < x^2 < 2\beta/3.
\end{cases}
\]
Set $\beta^* = \beta/(1 - \gamma)$, $x^* = x/\sqrt{1 - \gamma}$, $t^* = x^*/2 + \beta^*/x^*$. The condition $dE_{H_{0,1}} \Delta^2_1 = o(1)$ is equivalent to $\liminf_{d \to +\infty} v_d^* < 0$, where
\[
v_d^* = v_d/(1 - \gamma) = 1 + \begin{cases} -\beta^* & \text{as } (x^*)^2 \geq 2\beta^*, \\
-\beta^* + (x^*)^2 & \text{as } 2\beta^*/3 \leq (x^*)^2 \leq 2\beta^*, \\
\end{cases}
\]
This condition is equivalent to $\limsup_{d \to +\infty} x^* - \phi(\beta^*) < 0$ for $\beta^* \in (1/2, 1)$.

Yu.I.I. was partially supported by the RFBI grant 08-01-00692-a and by grant NSh–638.2008.1. A.B.T. would like to thank the Isaac Newton Institute for Mathematical Sciences in Cambridge for hospitality during the programme Statistical Theory and Methods for Complex, High-Dimensional Data. His research was partially supported by the grants EPSRC 531174, ANR-06-BLAN-0194 and by the PASCAL Network of Excellence. C.P. was partially supported by the grant ANR-07-BLAN-0234 and by PICS–2715.

References

Phil. Trans. R. Soc. A (2009)
Ingster, Yu. I. 1997 Some problems of hypothesis testing leading to infinitely divisible distributions. 
Methods Stat. 11, 37–68.
Ingster, Yu. I. & Suslina, I. A. 2002b On a detection of a signal of known shape in multichannel 
1723–1736.)
Ingster, Yu. I., Pouet, C. & Tsybakov, A. B. 2009 Sparse classification boundaries 
(doi:10.1214/009053607000000244)
Jin, J. 2004 Detecting a target in very noisy data from multiple looks. In A festschrift for 
Beachwood, OH: Institute of Mathematical Statistics.
recherches, Université Aix-Marseille 1.
Ross, M. E. et al. 2003 Classification of pediatric acute lymphoblastic leukemia by gene expression 
Ross, M. E. et al. 2004 Gene expression profile of pediatric acute myelogenous leukemia. Blood 104, 
Yeoh, E. J. et al. 2002 Classification, subtype discovery, and prediction of outcome in 
pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell 1, 133–143. 
(doi:10.1016/S1535-6108(02)00032-6)