Impossibility of successful classification when useful features are rare and weak

Jiashun Jin[1]

Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213

Communicated by David L. Donoho, Stanford University, Stanford, CA, April 9, 2009 (received for review February 21, 2009)

We study a two-class classification problem with a large number of features, out of which many are useless and only a few are useful, but we do not know which ones they are. The number of features is large compared with the number of training observations. Calibrating the model with 4 key parameters—the number of features, the size of the training sample, the fraction, and strength of useful features—we identify a region in parameter space where no trained classifier can reliably separate the two classes on fresh data. The complement of this region—where successful classification is possible—is also briefly discussed.

higher criticism | phase diagram | region of impossibility | region of possibility | threshold feature selection

An overwhelming trend in modern research activity is the tendency to gather very large databases and use them to search for good data-based classifier rules. For example, currently, a very large number of research teams in the medical sciences seek to gather and study gene expression microarray data in hopes of obtaining empirical rules that separate healthy patients from those affected by a disease—thus allowing for automatic diagnosis.

Much of the current surge of enthusiasm for such studies stems from the advent of high-thoughput methods that automatically, on each subject, make measurements of a very large numbers of features. In genomics, proteomics, and metabolomics it is now common to take several thousand automatic measurements per study subject. The opportunity to survey so many features at once is thought to be valuable: optimists will say that "surely somewhere among these many features will be a few useful ones allowing for successful classification!"

Advocates of the optimistic viewpoint must contend with the growing awareness in at least some fields that many published associations fail to replicate—i.e., the published classification rules simply do not work when applied to fresh data. Such failure has been the focus of meetings and special publications (1). Although there may be many reasons for failure to replicate (2, 3), we focus here on one specific cause: there may simply be too many useless features being produced by high-throughput devices, so that, even where there really are decisive features to be found in the high-throughput measurements, they simply cannot be reliably identified.

In fact, we establish in this article a specific "region of impossibility" for feature selection in classifier design. We identify settings with large numbers of measurements, some useful, some useless, where the subset of useful measurements, *if only it were known a priori*, would allow for training of a successful classifier; however, when the subset of useful features is not known, we show that no classifier-training procedure can be effective.

Specifically, we study a model problem introduced in refs. 4 and 5 where there are a large number of features, many of which are useless and a few of which are useful. In this model we consider a two-class classification problem where there are parameters controlling the fraction of useful features, the strength of the useful features, and the ratio between the number of observational units (e.g., patients) and the number of measured features (e.g., gene expression measurements). We identify a region in parameter space where, *with* prior knowledge of at least some useful features, success is possible, but *absent* such prior knowledge about the subset of useful features, no classifier built from the dataset is likely to separate the two classes on fresh data.

In companion work (5), we show that in the complement of this region, a specific method for classifier training—Higher Criticism Threshold feature selection (4)—does work, and so the results here are definitive.

Classification When Features Are Rare and Weak

Consider a two-class classification setting where we have a set of labeled training samples (Y_i, X_i), $i = 1, 2, \ldots, n$. Each label $Y_i = 1$ if X_i comes from class 1 and $Y_i = -1$ if X_i comes from class 2, and each feature vector $X_i \in R^p$. For simplicity, we suppose that the training set contains equal numbers of samples from each of the two classes, and that the feature vector obeys $X_i \sim N(Y_i\mu, I_p)$, $i = 1, 2, \ldots, n$, for an unknown mean contrast vector $\mu \in R^p$. Also, we suppose that the feature covariance matrix is the identity matrix. Extension to correlated cases is possible if side information about the feature covariance is available (see ref. 6, for example).

Following the two companion papers (4, 5), we consider the following *rare/weak feature model* (RW model), where the vector μ is nonzero in only an ϵ faction of coordinates, and the nonzero coordinates of μ share a common amplitude μ_0. Formally speaking, let $I_1, I_2, \ldots, I_p$ be samples from Bernoulli(ϵ), and let

$$\mu(j) = \mu_0 \cdot I_j, \quad 1 \le j \le p.$$

Let Z denote the vector of z scores corresponding to the training set: $Z(j) = (1/\sqrt{n}) \sum_{i=1}^{n} Y_i \cdot X_i(j)$. The j th z score arises in a formal normal-theory test of whether the j th feature is useless or useful. Under our assumptions, $Z \sim N(\sqrt{n}\mu, I_p)$; thus each coordinate of Z has expectation either 0 or $\tau = \sqrt{n}\mu_0$.

We assume $p \gg n$, ϵ is small, and τ is either small or moderately large (e.g., $p = 10{,}000$, $n = 100$, $\epsilon = 0.01$, $\tau = 2$). Because zero coordinates of μ are entirely noninformative for classification, the useful features are those with nonzero coordinates in μ. The parameters ϵ and τ can be set to make such useful features arbitrarily rare (by setting ϵ close to 0) and weak (setting τ small); we denote an instance of the rare/weak model by $RW(\epsilon, \tau; n, p)$.

Formally, our goal is to use the training data to design a classifier for use on fresh data. If we are given a new unlabeled feature vector X, we must then label it with a class prediction, i.e., attach a label $\hat{Y} = 1$ or $\hat{Y} = -1$. We hope that our predicted label $\hat{Y}$ is typically correct. The central problem is for which combinations (ϵ, τ, n, p) it is possible to train a classifier that can label Y correctly, and for which combinations is it not possible to do so?

Linking Rarity and Weakness to Number of Features. We now adopt an *asymptotic* viewpoint. We let the number of features p be the driving problem size descriptor, and for the purposes of calculation, we let p tend to infinity, and other quantities vary with p. We have checked that our asymptotic calculations are descriptive of actual classifier performance in realistic finite-sized problems, say

Author contributions: J.J. designed research, performed research, and wrote the paper.

The author declares no conflict of interest.

Freely available online through the PNAS open access option.

[1]E-mail: jiashun@stat.cmu.edu.

Table 1. Regimes and their definitions

Regime	Label	Definition
No growth	(N)	$n_p = n_0$ for some constant n_0
Slow growth	(S)	$n_p \to \infty$, but $n_p/p^\theta \to 0$, $\forall \theta > 0$
Regular growth	(R)	$n_p = p^\theta$ for some $\theta \in (0, 1)$

with p in the few thousands, as is now common in genomics and proteomics. Other problem parameters (fraction and strength of useful features, sample size n) will depend on p as follows. Fixing parameters $(\beta, r) \in (0,1)^2$, let

$$\epsilon = \epsilon_p = p^{-\beta}, \quad \tau = \tau_p = \sqrt{2r \log p}.$$

As $p \to \infty$, the useful features become increasingly rare; an asymptotically negligible fraction of the components in the vector Z. The parameters (β, r) describe the linkage between rareness and weakness of the entries in the parameter vector; they have been used before in classification studies (5) and more generally in detection studies (7–9). The domain $(\beta, r) \in (0,1)^2$ has been shown in earlier work to have an interesting two-phase structure; one can show there is a curve such that certain procedures succeed asymptotically when (β, r) lies above the curve and fail when (β, r) lies below the curve. We call a depiction of this domain and its phases a *phase diagram*. In this article, we exhibit a phase diagram such that, in the failure phase, every sequence of classification rules must fail for large p.

Linking Number of Observations to Number of Features. In classical statistical theory, one held p fixed and let n increase indefinitely. However, in modern scientific practice it seems the reverse is happening: one forms the impression that n stays fixed or grows very weakly while p grows dramatically (as high-throughput devices measure ever more features).

The phase diagram depends on the relationship between the number of features p and the number of study units n. Again, in our work it is convenient to make p the driving variable, and so $n = n_p$.

We can identify three regimes for the linkage between n and p: $n = n_p$ can have *no growth*, *slow growth*, or *regular growth*. Our labels for these regimes and their definitions are listed in Table 1.

The case of slow growth in our setting was previously studied in ref. 5; there, the focus was on the performance of a specific classifier-training procedure. Here, we study several types of linkages between n and p, and we also briefly discuss the case where n has an irregular growth, (see below). We are interested in limits that all classifier-training procedures must obey.

Asymptotic Rare/Weak Model (ARW). Combining the two linkages we have just discussed gives us the *asymptotic rare/weak model* $ARW(\beta, r, n_p)$. For each linkage type n_p we seek to identify ranges of (β, r) where successful classification is possible and impossible, respectively.

Impossibility of Classification. We will show that in each of the three growth regimes, there is a curve $r = \rho^\star(\beta)$ $(\star = N, S, R)$ which partitions the β-r plane into two components: a *region of impossibility* below the curve and and *region of possibility* above it. In detail, define the *standard phase boundary* function

$$\rho(\beta) = \begin{cases} 0, & 0 < \beta \le 1/2, \\ \beta - 1/2, & 1/2 < \beta < 3/4, \\ (1 - \sqrt{1-\beta})^2, & 3/4 \le \beta < 1. \end{cases} \quad [1]$$

The function ρ has appeared before in determining phase boundaries in a seemingly unrelated problem of multiple hypothesis testing (7–9). Define

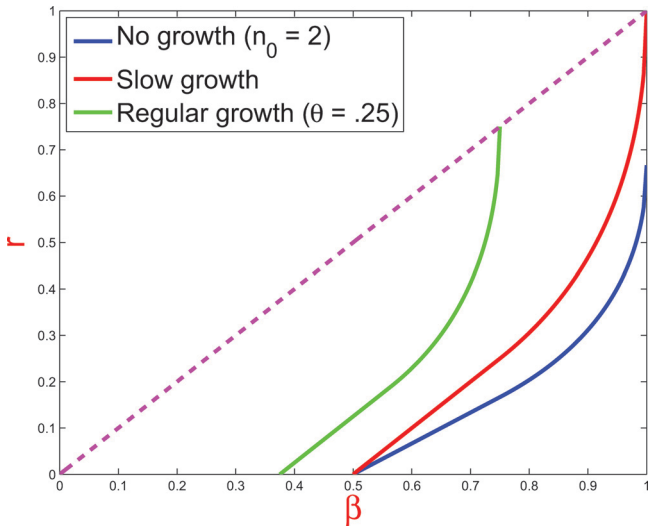

Fig. 1. Display of $r = \rho^\star(\beta)$, the boundary separating the region of impossibility and region of possibility, for the three types of growth of n: no growth (blue), slow growth (red), and regular growth (green). Also included is the diagonal line (magenta dashed) that illustrates the relationship in Eq. **2**.

$$\rho^N(\beta) = \rho^N(\beta, n) = \frac{n}{n+1}\rho(\beta), \quad 0 < \beta < 1,$$
$$\rho^S(\beta) = \rho(\beta), \quad 0 < \beta < 1,$$
$$\rho^R(\beta) = (1-\theta)\rho(\beta/(1-\theta)), \quad 0 < \beta < (1-\theta).$$

See Fig. 1. Note that in the definition of $\rho^R(\beta)$, we limit β to the range $(0, 1-\theta)$. Note also that for all three cases,

$$\rho^\star(\beta) \le \beta, \quad \star = N, S, R. \quad [2]$$

Definition 1: Fix $(\beta, r) \in (0,1)^2$. Fix one of the three types of growth of n by choosing $\star \in \{S, N, R\}$. We say (β, r) fall in the region of impossibility of the ARW if $r < \rho^\star(\beta)$.

Theorem 1. *Fix a growth regime n_p and fix a point (β, r) in the region "below" the corresponding graph $(\beta, \rho^\star(\beta))$. Consider the sequence of problems $ARW(r, \beta, n_p)$ for increasing p and a sequence of classifier-training methods, perhaps also dependent on p. The misclassification error rate of the resulting sequence of trained classifiers $\to 1/2$ with increasing p.*

In this region, the measurements are effectively noninformative, and random guessing does almost as well. However, note that there *are* useful features among the p features, and *if we only knew which features they were, we could reliably separate the classes!* Indeed, simply summing the coordinates known to be useful features and taking the signum would do the trick. In this sense, the region of impossibility is precisely the region where the effect mentioned in the introduction shows up; *the attempt to find the useful features among many useless ones is simply doomed.* Note that Fan and Fan (10) studied a closely related setting and identified a different region of impossibility. See details therein.

Conversely, one can show that fixing (β, r) in the region of possibility, successful classifier training *is* possible, and there is a sequence of trained classifiers whose misclassification probability $\to 0$ as $p \to \infty$. However, that is beyond the scope of this research announcement; we refer the reader to the author's related papers. See Fig. 2 for a display of region of impossibility and region of possibility.

Proof of Theorem 1: To understand the role of the training data, we compare the problem of classification *with* Z and that *without* Z. When Z is not available, the test features X contains $\approx p\epsilon_p$ useful features, each of which has a strength of $\pm\tau_p/\sqrt{n}$ (with sign "+" if

X is from class 1 and "−" otherwise). In this case, the classification problem reduces to the testing problem studied in our previous work (8), and it is possible to successfully classify if and only if $r/\sqrt{n} > \rho(\beta)$ (7, 8).

When Z is available to us, the picture is very different. For any $1 \le i \le p$, the probability that the ith coordinate of X contains a useful feature is no longer ϵ_p, but instead the posterior probability $\eta_i = \eta(Z_i;p)$, where

$$\eta(z;p) = \epsilon_p\varphi(z-\tau_p)/[(1-\epsilon_p)\varphi(z)+\epsilon_p\varphi(z-\tau_p)], \qquad [3]$$

with φ being the density of $N(0,1)$. This is a monotone function, ≈ 0 for small z and ≈ 1 for large z. Intuitively, a large coordinate amplitude in Z suggests a useful feature, and a moderate or small amplitude suggests a useless one. Seemingly, this is a different model from the previous case, the study of which needs different analytic technique.

First, denote the density of $N(0,I_p)$ by $f_0^{(p)} = f_0^{(p)}(x_1,x_2,\ldots,x_p)$. Second, for $k = 1,2$, denote the conditional density of $(X|Z)$ when $X \sim$ Class k by

$$f_k^{(p)} = f_k^{(p)}(x_1,x_2,\ldots,x_p|Z;\epsilon_p,\tau_p,n_p,p),$$

and denote the conditional density of $(X_1|Z_1)$ by

$$f_k^{(1)} = f_k^{(1)}(x_1|Z_1;\epsilon_p,\tau_p,n_p,p).$$

Here, X_1 and Z_1 are the first coordinates of X and Z, respectively. Finally, for two density functions f and g, define the Hellinger affinity by $H(f,g) = \int\sqrt{f(x)g(x)}dx$. Let the (conditional)-Hellinger affinity between $f_0^{(p)}$ and $f_1^{(p)}$ be $H(f_0^{(p)},f_1^{(p)};Z,\epsilon_p,\tau_p,n_p,p)$, and that between $f_0^{(1)}$ and $f_1^{(1)}$ be $H(f_0^{(1)},f_1^{(1)};Z_1,\epsilon_p,\tau_p,n_p,p)$. We have the following lemma.

Lemma 1. *Fix n_p and $(\beta,r) \in (0,1)^2$ in the ARW(β,r,n_p) model. For any classifier $T = T(X,Z;p)$*

$$|P\{misclassification|T\} - 1/2| \le C\big(1-E[H(f_0^{(p)},f_1^{(p)};Z,\epsilon_p,\tau_p,n_p,p)]\big)^{1/2}.$$

We omit the proof of *Lemma 1*. Relationships between classification error rate and Hellinger affinity are well known. In this case, the added wrinkle is to condition on the training data Z. Besides that feature, the argument is standard; see ref. 11 for example.

The following lemma is elementary; we omit the proof.

Lemma 2. $E[H(f_0^{(p)},f_1^{(p)};Z,\epsilon_p,\tau_p,n_p,p)] = (E[H(f_0^{(1)},f_1^{(1)};Z_1,\epsilon_p,\tau_p,n_p,p)])^p$.

The heart of the proof of *Theorem 1* is the following lemma. Its proof is relatively long, so we leave it to later sections.

Lemma 3. *Fix one of the three growth types; for any fixed parameters (β,r) in the corresponding region of impossibility of the ARW(β,r,n_p), $E[H(f_0^{(1)},f_1^{(1)};Z_1,\epsilon_p,\tau_p,n_p,p)] = 1 + o(1/p)$, $p \to \infty$.*

Combining these lemmas gives *Theorem 1*.

Extension to Cases Where *n* Grows Irregularly. So far, we considered n_p growing with p according to one of three specific regimes: (N), (S), (R). However, the conclusion of *Theorem 1* can be obtained in a much broader range of cases, where n grows somewhat irregularly.

Lemma 4 below says that $E[H(f_0^{(p)},f_1^{(p)};Z,\epsilon,\tau,n,p)]$ is a monotone function of the sample size n. This implies that, if n_p is *eventually sandwiched* between two sequences obeying one of our growth regimes, then its behavior is also sandwiched between the results at those two regular situations.

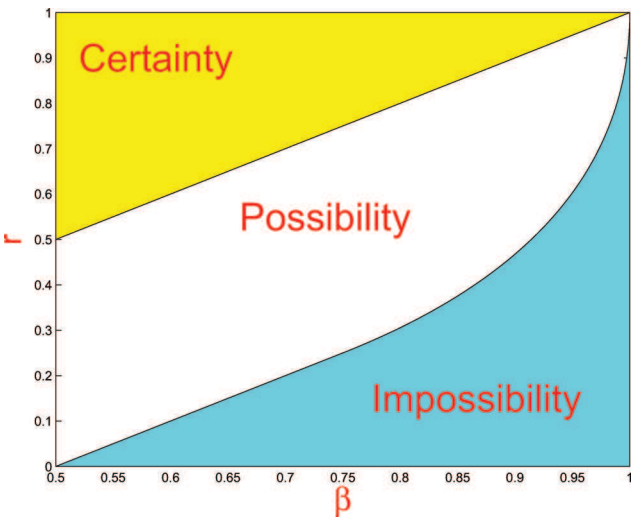

Fig. 2. Display of region of impossibility (cyran) and region of possibility (white and yellow) in the case of slow growth [only the range of $\beta \in (1/2,1)$ is shown]. In the Certainty region, it is not only possible to have successful classification, but also possible to identify nearly all useful features.

Lemma 4. *Fixing $p,\tau > 0$ and $\epsilon \in (0,1)$ in the RW$(\epsilon,\tau;n,p)$, $E[H(f_0^{(p)},f_1^{(p)};Z_1,\epsilon,\tau,n,p)]$ is a monotone increasing function of n.*

Proof: By *Lemma 2*, it suffices to show that $E[H(f_0^{(1)},f_1^{(1)};Z_1,\epsilon,\tau,n,p)]$ is a monotone increasing function of n. Note that $f_0^{(1)}$ is the density of $N(0,1)$, and that $f_1^{(1)}(x) = (1-\eta(Z_1))\varphi(x)+\eta(Z_1)\varphi(x-\mu_p)$, with $\eta(Z_1) = \eta(Z_1;p)$ being defined in Eq. **3**, and $\mu_p = \tau_p/\sqrt{n}$. By direct calculations,

$$H(f_0^{(1)},f_1^{(1)};Z_1) = E_0\left[\left(1-\eta(Z_1)+\eta(Z_1)e^{\mu_p X_1-\frac{\mu_p^2}{2}}\right)^{1/2}\right], \qquad [4]$$

where we have suppressed parameter dependencies and E_0 denotes the expectation with respect to the law of $X \sim N(0,1)$. Observe that $H(f_0^{(1)},f_1^{(1)};Z_1,\epsilon,\tau,n,p)$ depends on n only through μ_p, and that μ_p is monotone decreasing in n. It is sufficient to show that for any numbers $\eta \in (0,1)$ and $\mu' \ge \mu$,

$$E_0\left[\sqrt{(1-\eta)+\eta e^{\mu' X-\frac{\mu'^2}{2}}}\right] \le E_0\left[\sqrt{(1-\eta)+\eta e^{\mu X-\frac{\mu^2}{2}}}\right]. \qquad [5]$$

Toward this end, write $\mu' X_1 = \mu U + \delta W$, where U and W are iid samples from $N(0,1)$ and $\delta^2 = (\mu')^2 - \mu^2$. Inserting this into the left-hand side of Eq. **5** gives

$$E_0\left[\sqrt{(1-\eta)+\eta\cdot e^{\mu' X_1-(\mu')^2/2}}\right] = E_0\left[\sqrt{(1-\eta)+\eta\cdot e^{\mu U-\mu^2/2}e^{\delta W-\delta^2/2}}\right].$$

It follows from Jensen's inequality that

$$E_0\left[\sqrt{(1-\eta)+\eta e^{\mu' X-\frac{\mu'^2}{2}}}\right] \le E_0\left[\sqrt{(1-\eta)+\eta\cdot e^{\mu U-\mu^2/2}E[e^{\delta W-\delta^2/2}]}\right]$$

where the right-hand side reduces to that of Eq. **5**. This gives the claim.

Gaussian Assumption. We now discuss the Gaussian assumption on Z (that on X_i has a less important effect). When n is relatively large (e.g., $n \geq 20$), the assumption is reasonable. When n is very small, it might be better to assume that $Z(j)$ are t-distributed. Because the tail of the t distribution is heavier than that of the Gaussian, it is harder to successfully classify in the t-error model than in the Gaussian model, given that the parameters (p, n, ϵ, τ) are the same in two settings. Therefore, the region of impossibility continues to be valid in the t setting. Of course, the exact boundary separating region of possibility and region of impossibility depends on the specific tail behavior of the marginal density of $Z(j)$, and may be different from that in this note.

Numerical Examples. In companion articles (4, 5), we show that threshold feature selection has an optimal region of classification performance for the ARW model, when the threshold t is chosen ideally. Such a classifier has the form $L(X;t) = \sum_{j=1}^{p} \mathrm{sgn}(Z_j) 1_{\{|Z_j| \geq t\}} X(j)$. Therefore, we can use $L(X;t)$ with an ideal choice of t to investigate the minimum misclassification error that one can achieve. Call the minimum error the ideal error and the minimizing t the ideal threshold; both quantities can be conveniently calculated assuming that we know (ϵ, τ).

Fix $p = 10^5$, let $n = 2$, 10, and 100, representing the three types of growth (N), (S), and (R) [the parameter $\theta = 0.4$ in (R)]. When $n = 2, 10$, let $\beta = 0.55, 0.65, 0.75$ and $\epsilon = p^{-\beta} \approx 178/56/18 \times 10^{-5}$. When $n = 100$, let $\beta = 0.35/0.45/0.55$ and $\epsilon = p^{-\beta} \approx 1,778/562/178 \times 10^{-5}$. For each triplet (p, n, ϵ), let τ range from 0.5 to 3 with an increment of 0.1, and calculate the ideal error. Define $\tau^{\diamond} = \tau^{\diamond}(p, n, \epsilon)$ as the largest τ such that the ideal error $\geq 40\%$ (say); this can be thought of as the critical value below which successful classification is quasi-impossible. (Note: the choice of 40% for the critical value is arbitrary; other choices produce quantitatively similar but not identical results). When $n = 2$ (and β takes corresponding values as above), $\tau^{\diamond} \approx 0.9/1.5/2.0$; when $n = 10$, $\tau^{\diamond} \approx 1.3/1.9/2.4$; when $n = 100$, $\tau^{\diamond} \approx 0.7/1.3/1.9$.

We compare $\tau^{\diamond}$ with the asymptotic critical value $\tau^{\star} \equiv \sqrt{2\rho^{\star}(\beta)\log p}$ as in *Theorem 1*. Given $p = 10^5$, when $n = 2$ (and β takes corresponding values as above), $\tau^{\star} \approx 0.88/1.5/1.96$; when $n = 10$, $\tau^{\star} \approx 1.07/1.86/2.40$; when $n = 100$, $\tau^{\star} \approx 1.07/1.86/2.64$. Both critical values—$\tau^{\diamond}$ and $\tau^{\star}$—are close to each other, especially when $n = 2, 10$. When $n = 100$, the differences between two critical values are large but still get smaller for larger p. This suggests that the asymptotic separating boundary $r = \rho^{\star}(\beta)$ is valid already for $p = 10^5$.

Last, for each combination (p, n, ϵ, τ), we calculate the ideal threshold $t^* = t^*(p, n, \epsilon, \tau)$, and apply $L(X; t^*)$ to samples generated according to $RW(p, n, \epsilon, \tau)$. In Table 2, we report the average (empirical)-misclassification errors across 1,000 independent repetitions. (To save space, only part of the results are reported.) Cells in boldface/nonboldface correspond to τ's that fall below/above $\tau^{\star}$, respectively. As predicted in *Theorem 1*, most boldface numbers $\geq$40%, and most nonboldface numbers <40% and get increasingly smaller as τ increases. Also, the results suggest that in the region of impossibility, $L(X;t)$ performs poorly even with ideal threshold.

Relation to Higher Criticism. We briefly discuss the region of possibility. In the interior of this region, it is possible to train classifiers whose misclassification probability on fresh data $\to 0$ under the ARW model. Such a classifier can be trained by adopting the recent notion of Higher Criticism (HC) Threshold feature selection.

HC was first introduced in our previous work (8) as follows. Given a collection $\pi_{(1)}, \ldots, \pi_{(p)}$ of sorted P values, one calculates the HC objective values

Table 2. Misclassification errors for $L(X;t)$ with ideal thresholds ($p = 10^5, \epsilon = p^{-\beta}$)

τ	$n = 2$ (N) $\beta = 0.55/0.65/0.75$	$n = 10$ (S) $\beta = 0.55/0.65/0.75$	$n = 100$ (R), $\theta = 0.4$ $\beta = 0.35/0.45/0.55$
0.6	**0.406/0.480/0.492**	**0.501/0.512/0.473**	**0.411/0.464/0.503**
0.9	**0.387/0.466/0.470**	**0.441/0.474/0.508**	**0.354/0.466/0.477**
1.2	0.285/**0.423/0.456**	0.434/**0.443/0.477**	0.249/**0.388/0.485**
1.5	0.197/**0.376/0.474**	0.358/**0.435/0.487**	0.137/**0.354/0.420**
1.8	0.067/0.329/**0.456**	0.245/**0.419/0.490**	0.028/**0.283/0.414**
2.1	0.002/0.170/0.412	0.109/0.341/**0.428**	0.002/0.139/**0.348**
2.4	0.000/0.054/0.315	0.016/0.232/**0.392**	0.000/0.018/**0.276**
2.7	0.000/0.013/0.180	0.001/0.088/0.350	0.000/0.001/0.151
3.0	0.000/0.000/0.082	0.000/0.017/0.220	0.000/0.000/0.055

Cells in **boldface** correspond to τ's that fall below the critical value $\tau^{\star} = \sqrt{2\rho^{\star}(\beta)\log p}$. These exhibit high classification errors, as predicted. The strength of useful features is $\tau/\sqrt{n}$, so the classification problem is increasingly harder for larger n.

$$HCObj(k) = \sqrt{p}\frac{\pi_{(k)} - k/p}{\sqrt{k/p(1-k/p)}}, \quad k = 1, \ldots, p.$$

The HC statistic is the maximum of the objective function. It can be used to assess significance of the whole body of P values. Given the feature vector X as in the ARW (but not any class labels), test whether the mean vector $\mu = 0$ identically, or whether μ contains a fraction $\epsilon_p > 0$ of nonzero coordinates, each of them equal to an unknown parameter τ_p. The testing problem is a modification of the classification problem we study in this note, where the training set is not available. Similarly, the testing problem was shown in refs. 7 and 8 to have a phase diagram (β, r) with a region of impossibility and a region of possibility. In fact, if we express $\tau_p = \sqrt{2r'\log p}$ and $\epsilon_p = n^{-\beta}$, then the region of impossibility is the range of (β, r) that satisfy $r' < \rho(\beta)$ and $0 < \beta < 1$, where ρ is the standard phase boundary function introduced in Eq. **1** and $r' = r'(r, \star)$ is the calibration of r appropriate to growth regime $\star \in \{N, S, R\}$. The region of possibility is $r' > \rho(\beta)$ and $0 < \beta < 1$. In the whole region of possibility, HC was shown in ref. 8 to yield a successful test: the sum of type I and type II errors of the test $\to 0$ as $p \to 0$. See ref. 8 (and also refs. 7 and 9) for details.

HC can be used to select thresholds for feature selection (4). One maximizes the HC objective over the interval $1 \leq k \leq \alpha_0 \cdot p$. The P value $\pi_{(k^*)}$ at the maxmizer can be converted into a two-sided Z score, say $z_{(k^*)}$. Select all features whose feature Z scores exceed $z_{(k^*)}$ in absolute value. The trained classifier is the weighted sum of the standardized feature values, with weights obtained by thresholding the training set Z scores at the HCT $z_{(k^*)}$. The concept, numerical performance, and practical features of HCT are reported in ref. 4, and an idealized HCT was carefully studied in ref. 5 in the slow-growth regime (S). It was shown that in the slow-growth regime (S), ideal HCT works throughout the possibility region of the phase diagram. The idea of component-wise thresholding is closely related to that in ref. 10 (see also refs. 12 and 13 where the focus is on hypothesis testing and dimension reduction, respectively).

HCT Achieves Separation Throughout the Possibility Region. First, we show that if we perform ideal feature selection with an oracle threshold—i.e., if an oracle tells us the unknown parameters (β, r)—then the resulting trained classifier yields successful classification throughout the region of possibility. Second, we show that that the HCT converges to the oracle threshold asymptotically (but does not require help from any oracle).

HC can also been used directly for classification (14) without feature selection (for comparison with the method above, see ref. 5).

In this article, we have attempted to draw the attention of working scientists to a basic phenomenon that might affect many high-throughput studies. The mathematical scientist interested in this phenomenon will want to know that independently, Ingster, Pouet, and Tsybakov (15) have analyzed a setting more general than the present one and identified similar phase transitions phenomena.

Proof of Lemma 3: We now show *Lemma 3* for the case of no growth and the case of regular growth. Once these are proved, the case of slow growth follows by the monotonicity result in *Lemma 4* and the way $\rho^\star(\beta)$ is defined.

The case of no growth (N). In this case, n is a fixed integer. It suffices to show for fixed (β, r) with $r < \frac{n+1}{n}\rho(\beta)$,

$$E\big[H\big(f_0^{(1)}, f_1^{(1)}; Z_1\big)\big] = 1 + o(1/p). \qquad [6]$$

Note that in the ARW, the density of Z_1 is $(1-\epsilon_p)\varphi(x)+\epsilon_p\varphi(x-\tau_p)$. By Eq. **4** and direct calculations,

$$E\big[H\big(f_0^{(1)}, f_1^{(1)}; Z_1\big)\big] = (1-\epsilon_p)I + \epsilon_p II, \qquad [7]$$

where

$$I = E_{0,0}\left(\frac{(1-\epsilon_p)+\epsilon_p e^{\tau_p Z-\tau_p^2/2}e^{\mu_p X-\mu_p^2/2}}{(1-\epsilon_p)+\epsilon_p e^{\tau_p Z-\tau_p^2/2}}\right)^{1/2},$$

$$II = E_{0,0}\left(\frac{(1-\epsilon_p)+\epsilon_p e^{\tau_p Z+\tau_p^2/2}e^{\mu_p X-\mu_p^2/2}}{(1-\epsilon_p)+\epsilon_p e^{\tau_p Z+\tau_p^2/2}}\right)^{1/2},$$

and $E_{0,0}$ is the expectation with respect to the law that X and Z are iid samples from $N(0,1)$. Write for short $E_0 = E_{0,0}$ whenever there is no confusion. Introduce $a_p = 1/(1-\epsilon_p)$, $V_1(\theta,\zeta) = (\frac{1+a_p\theta\zeta}{1+a_p\theta})^{1/2} - 1 - \frac{1}{2}a_p\theta\zeta + \frac{1}{2}a_p\theta$, and $V_2(\theta,\zeta) = (\frac{1+a_p\theta\zeta}{1+a_p\theta})^{1/2} - 1$. Note that

$$E_0\big[e^{\tau_p Z+\mu_p X-\tau_p^2/2-\mu_p^2/2}\big] = 1, \quad E_0\big[e^{\mu_p X-\mu_p^2/2}\big] = 1. \qquad [8]$$

It follows from direct calculations that

$$\begin{aligned} I &= 1 + \frac{a_p\epsilon_p}{2}\big(E_0\big[e^{\tau_p Z+\mu_p X-\tau_p^2/2-\mu_p^2/2}\big] - E_0\big[e^{\tau_p Z-\tau_p^2/2}\big]\big) \\ &\quad + E\big[V_1\big(\epsilon_p e^{\tau_p Z-\tau_p^2/2}, e^{\mu_p X-\mu_p^2/2}\big)\big] \\ &= 1 + E_0\big[V_1\big(\epsilon_p e^{\tau_p Z-\tau_p^2/2}, e^{\mu_p X-\mu_p^2/2}\big)\big], \end{aligned} \qquad [9]$$

and, similarly,

$$II = 1 + E_0\big[V_2\big(\epsilon_p e^{\tau_p Z-\tau_p^2/2}, e^{\mu_p X-\mu_p^2/2}\big)\big]. \qquad [10]$$

Combine Eqs. **7–10**; to show Eq. **6**, it is sufficient to show

$$E_0\big[V_1\big(\epsilon_p e^{\tau_p Z-\tau_p^2/2}, e^{\mu_p X-\mu_p^2/2}\big)\big] = o(1/p), \qquad [11]$$

and

$$E_0\big[V_2\big(\epsilon_p e^{\tau_p Z-\tau_p^2/2}, e^{\mu_p X-\mu_p^2/2}\big)\big] = o(p^{\beta-1}). \qquad [12]$$

Toward this end, introduce

$$\psi_1(x) = \begin{cases} 2x-1, & x > 1, \\ x^2, & x \le 1, \end{cases} \qquad \psi_2(x) = \min\{x, 1\}.$$

We need the following lemma, the proof of which is elementary so we omit it.

Lemma 5. *For sufficiently large p, there is constant $C > 0$ such that for any $\theta > 0$ and $\zeta > 0$, $|V_1(\theta,\zeta)| \le C[\psi_1(\theta\zeta) + (1+\zeta)\psi_1(\theta)]$ and $|V_2(\theta,\zeta)| \le C(1+\zeta)\psi_2(\theta)$.*

We now show Eqs. **11** and **12**. Consider Eq. **11** first. Denote

$$\sigma_p = \big(\tau_p^2 + \mu_p^2\big)^{1/2}, \quad W = (\tau_p Z + \mu_p X)/\sigma_p. \qquad [13]$$

As X and Z are iid samples from $N(0,1)$, so $W \sim N(0,1)$. Using *Lemma 5*,

$$\begin{aligned} &\big|E_0\big[V_1\big(\epsilon_p e^{\tau_p Z-\tau_p^2/2}, e^{\mu_p X-\mu_p^2/2}\big)\big]\big| \qquad [14] \\ &\le C\big[E_0\big[\psi_1\big(\epsilon_p e^{\tau_p Z+\mu_p X-\tau_p^2/2-\mu_p^2/2}\big)\big] \\ &\quad + E_0\big[\big(1+e^{\mu_p X-\mu_p^2/2}\big)\psi_1\big(\epsilon_p e^{\tau_p Z-\tau_p^2/2}\big)\big]\big] \\ &= C\left[E_0\left[\psi_1\left(\epsilon_p e^{\sigma_p W-\frac{\sigma_p^2}{2}}\right)\right] + E_0\left[\psi_1\left(\epsilon_p e^{\tau_p Z-\frac{\tau_p^2}{2}}\right)\right]\right]. \qquad [15] \end{aligned}$$

The second term in Eq. **15** is no greater than the first term. To see this, writing $\psi_1(\epsilon_p e^{\sigma_p W-\sigma_p{}^2/2}) = \psi_1(\epsilon_p e^{\tau_p Z-\tau_p^2/2}e^{\mu_p X-\mu_p^2/2})$, it follows from Eq. **8**, the convexity of ψ_1, and Jensen's inequality, that

$$\begin{aligned} E_0\left[\psi_1\left(\epsilon_p e^{\tau_p Z-\tau_p^2/2}\right)\right] &= E_0\left[\psi_1\left(\epsilon_p e^{\tau_p Z-\frac{\tau_p^2}{2}}E_0\left[e^{\mu_p X-\frac{\mu_p^2}{2}}\right]\right)\right] \\ &\le E_0\left[\psi_1\left(\epsilon_p e^{\sigma_p W-\sigma_p^2/2}\right)\right], \end{aligned}$$

which validates the aforementioned point. Combining this with Eq. **15** gives

$$\left|E_0\left[V_1\left(\epsilon_p e^{\tau_p Z-\frac{\tau_p^2}{2}}, e^{\mu_p X-\frac{\mu_p^2}{2}}\right)\right]\right| \le CE_0\left[\psi_1\left(\epsilon_p e^{\sigma_p W-\frac{\sigma_p{}^2}{2}}\right)\right]. \qquad [16]$$

We now analyze $E_0[\psi_1(\epsilon_p e^{\sigma_p W-\frac{\sigma_p{}^2}{2}})]$. Introduce $r_0 = \frac{n+1}{n}r$. Recall that $\tau_p^2 = 2r\log p$ and that $\mu_p^2 = (2/n)r\log p$. It follows from the definition of σ_p and r_0 that $\sigma_p = \sqrt{2r_0\log p}$. In addition, we introduce two thresholds $t_p = t_p(r,\beta)$ and $t_p^0 = t_p^0(r_0,\beta)$ by

$$t_p = \frac{\beta+r}{2r}\cdot\tau_p, \quad t_p^0 = \frac{\beta+r_0}{2r_0}\cdot\sigma_p. \qquad [17]$$

By these definitions and basic algebra,

$$\epsilon_p e^{\sigma_p W-\sigma_p{}^2/2} > 1 \text{ if and only if } W > t_p^0 \qquad [18]$$

Combining Eq. **18** with the definition of ψ_1,

$$\begin{aligned} &E_0\big[\psi_1\big(\epsilon_p e^{\sigma_p W-\sigma_p{}^2/2}\big)\big] \\ &\le 2\int_{t_p^0}^{\infty}\epsilon_p e^{\sigma_p w-\sigma_p{}^2/2}\varphi(w)dw + \int_{-\infty}^{t_p^0}\epsilon_p^2 e^{2\sigma_p w-\sigma_p{}^2}\varphi(w)dw \\ &= 2\epsilon_p\Phi\big(-\big(t_p^0-\sigma_p\big)\big) + \epsilon_p^2 e^{\sigma_p{}^2}\Phi\big(t_p^0-2\sigma_p\big), \end{aligned} \qquad [19]$$

where Φ is the cdf of $N(0,1)$.

In addition, by the assumption $r < \frac{n}{n+1}\rho(\beta)$, we have $r_0 < \rho(\beta)$. In view of the definitions of t_p, σ_p, and $\rho(\beta)$, it follows from basic algebra that

$$1 < \frac{t_p}{\sigma_p} \le 2 \text{ if } r_0 \ge \beta/3; \quad \frac{t_p}{\sigma_p} > 2 \text{ if } r_0 > \beta/3. \qquad [20]$$

Note also that

$$\Phi(t) \le C\varphi(|t|) \text{ for } t \le 0 \text{ and } \Phi(t) \le 1 \text{ for all } t. \qquad [21]$$

Combining Eqs. **19–21** gives

$$E_0\left[\psi_1\left(\epsilon_p e^{\sigma_p W-\sigma_p{}^2/2}\right)\right] \le \begin{cases} Cp^{-\frac{(\beta+r_0)^2}{4r_0}}, & \text{if } r_0 \ge \beta/3, \\ C\left[p^{-\frac{(\beta+r_0)^2}{4r_0}} + p^{-(2\beta-2r_0)}\right], & \text{if } r_0 < \beta/3. \end{cases}$$

STATISTICS

Recall that $r_0 < \rho(\beta)$. It follows from the definition of $\rho(\beta)$ that $\frac{(\beta+r_0)^2}{4r_0} > 1$, and that $2\beta - 2r_0 > 1$. As a result,

$$E_0\big[\psi_1\big(\epsilon_p e^{\sigma_p W - \sigma_p{}^2/2}\big)\big] = o(1/p). \quad \textbf{[22]}$$

Inserting Eq. **22** into Eq. **16** gives Eq. **11**.

Next, consider Eq. **12**. Similarly, by *Lemma 5*, Eq. **8**, and that X and Z are independent,

$$\left|E_0\left[V_2\left(\epsilon_p e^{\tau_p Z+\tau_p^2/2}, e^{\mu_p X-\mu_p^2/2}\right)\right]\right| \le CE_0\big[\big(1+e^{\mu_p X-\mu_p^2/2}\big) \times \psi_2\big(\epsilon_p e^{\tau_p Z+\tau_p^2/2}\big)\big] = CE_0\big[\psi_2\big(\epsilon_p e^{\tau_p Z+\tau_p^2/2}\big)\big]. \quad \textbf{[23]}$$

Similarly, as that $\epsilon_p e^{\tau_p Z+\tau_p^2/2} > 1$ if and only if $Z > t_p - \tau_p$, by the definition of ψ_2 and elementary calculus,

$$E_0\left[\psi_2\left(\epsilon_p e^{\tau_p Z+\tau_p^2/2}\right)\right] \quad \textbf{[24]}$$

$$\le \int_{t_p-\tau_p}^{\infty} \varphi(z)dz + \int_{-\infty}^{t_p-\tau_p} \epsilon_p e^{\tau_p z+\tau_p^2/2}\varphi(z)dz$$

$$= \Phi(-(t_p-\tau_p)) + \epsilon_p e^{\tau_p^2}\Phi(t_p-2\tau_p). \quad \textbf{[25]}$$

Combine Eq. **25** with Eqs. **20** and **21**,

$$E_0\left[\psi_2\left(\epsilon_p e^{\tau_p Z+\tau_p^2/2}\right)\right] \le \begin{cases} Cp^{-\frac{(\beta-r)^2}{4r}}, & r \ge \beta/3, \\ C\left[p^{-\frac{(\beta-r)^2}{4r}} + p^{-(\beta-2r)}\right], & r < \beta/3. \end{cases} \quad \textbf{[26]}$$

Now, since $r < \rho(\beta)$, then $\frac{(\beta-r)^2}{4r} > (1-\beta)$ and $(\beta - 2r) > 1-\beta$. Inserting these into Eq. **26** gives Eq. **12**, and concludes the proof for the case of no growth.

The case of regular growth (R). By Eq. **2**, we can limit (β, r) to the range of $0 < r < \beta$ and $0 < r < 1$. Define

$$\delta(r,\beta) = \begin{cases} r - \left(\beta - \frac{1}{2}\right), & \text{if } r \le \beta/3, \\ r - \left(\beta - \frac{1}{2} - \frac{(\beta-3r)^2}{8r}\right), & \text{if } r > \beta/3. \end{cases} \quad \textbf{[27]}$$

Basic algebra shows that the assumption $r < (1-\theta)\rho(\beta/(1-\theta))$ is equivalent to $2\delta(r,\beta) < \theta$. Recall that the sample size $n = p^\theta$. It is sufficient to show that for fixed (β, r, θ) satisfying $2\delta(r,\beta) < \theta$,

$$E[H(f_0,f_1;Z)] = 1 + o(1/p). \quad \textbf{[28]}$$

Rewrite $H(f_0,f_1;Z) = E_0[(1+\eta_p(Z)(e^{\mu_p X-\mu_p^2/2}-1))^{1/2}]$. Since that $|\sqrt{1+x} - 1 - x/2| \le Cx^2$ for any $x > -1$. then

$$\left|H(f_0,f_1;Z) - 1 - \frac{1}{2}\eta_p(Z)E_0\left[e^{\mu_p X-\mu_p^2/2} - 1\right]\right| \le C\eta_p^2(Z)E_0\left[\left(e^{\mu_p X-\mu_p^2/2}-1\right)^2\right]. \quad \textbf{[29]}$$

Recalling $n = p^\theta$ and $\mu_p = \tau_p/\sqrt{n}$, direct calculations show that

$$E_0\big(e^{\mu_p X-\mu_p^2/2} - 1\big) = 0, \quad \textbf{[30]}$$

$$E_0\big(e^{\mu_p X-\mu_p^2/2} - 1\big)^2 = e^{\mu_p^2} - 1 \le C\log(p)p^{-\theta}. \quad \textbf{[31]}$$

Combining Eqs. **30** and **31** with Eq. **29** gives $\big|H(f_0,f_1;Z) - 1\big| \le C\log(p)p^{-\theta}\eta_p^2(Z)$. If we can show that for any (β, r) in the range of $0 < r < \beta$ and $0 < \beta < 1$,

$$E\big[\eta_p^2(Z)\big] \le Cp^{-2+2\delta(r,\beta)}, \quad \textbf{[32]}$$

then $|E[H(f_0,f_1)] - 1| \le E[|H(f_0,f_1) - 1|] \le C\log(p)p^{-1+2\delta(r,\beta)-\theta}$, and Eq. **28** follows from the assumption of $2\delta(r,\beta) < \theta$.

We now show Eq. **32**. Write $E[\eta^2(Z)] = I + II$, where

$$I = (1-\epsilon_p)E_0[\eta^2(Z)], \quad II = \epsilon_p E_0[\eta^2(\tau_p + Z)], \quad \textbf{[33]}$$

with E_0 being the expectation with respect to the law of $Z \sim N(0,1)$. It is sufficient to show that

$$I \le Cp^{-1+2\delta(r,\beta)}, \quad II \le Cp^{-1+2\delta(r,\beta)}. \quad \textbf{[34]}$$

Consider the first claim of Eq. **34**. Recall that $\tau_p = \sqrt{2r\log p}$ and $t_p = t_p(r,\beta) = \frac{\beta+r}{2r}\tau_p$, and note that for sufficiently large p, $(1-\epsilon_p) + \epsilon_p e^{\tau_p Z-\tau_p^2/2} \ge 1/2$. Combine this with the way that $\eta(Z)$ is defined, $\eta(Z) \le 2\epsilon_p e^{\tau_p Z-\tau_p^2/2}$ when $Z \le t_p$ and $\eta(Z) \le 1$ otherwise. It follows from these inequalities and elementary calculus that

$$I \le C\left[\epsilon_p^2 e^{\tau_p^2}\Phi(t_p - 2\tau_p) + \Phi(-t_p)\right]. \quad \textbf{[35]}$$

By arguments similar to that in the proof for the case of no growth,

$$\Phi(-t_p) \le C\varphi(t_p) \le Cp^{-\frac{(\beta+r)^2}{4r}}, \quad \textbf{[36]}$$

and

$$\epsilon_p^2 e^{\tau_p^2}\Phi(t_p - 2\tau_p) \le \begin{cases} \epsilon_p^2\tau_p^2 = n^{2r-2\beta}, & \text{if } r \le \beta/3, \\ Cn^{2r-2\beta-\frac{(\beta-3r)^2}{4r}}, & \text{if } r > \beta/3. \end{cases} \quad \textbf{[37]}$$

Inserting Eqs. **36** and **37** into Eq. **35** gives the claim.

Consider the second claim of Eq. **34**. By similar argument, $II \le C[\epsilon_p^3 e^{3\tau_p^2}\Phi(t_p - 3\tau_p) + \epsilon_p\Phi(-(t_p - \tau_p))]$,

$$\epsilon_p^3 e^{3\tau_p^2}\Phi(t_p - 3\tau_p) \le \begin{cases} n^{6r-3\beta}, & r \le \beta/5, \\ Cn^{2r-2\beta-\frac{(\beta-3r)^2}{4r}}, & r > \beta/5, \end{cases}$$

and $\epsilon_p\Phi(-(t_p - \tau_p)) \le Cn^{2r-2\beta-\frac{(\beta-3r)^2}{4r}}$. Since $(6r - 3\beta) < (2r - 2\beta)$ when $r \le \beta/5$, combining these results gives the second claim of Eq. **34**, and concludes the proof for the case of regular growth.

ACKNOWLEDGMENTS. I thank David Donoho and anonymous referees for numerous helpful pointers and comments in improving the manuscript, and the Newton Institute for hospitality during the program Statistical Challenges of High-Dimensional Data. This work was supported in part by National Science Foundation Grant DMS-0908613.

1. NCI-NHGRI Working Group on replication in association studies (2007) Replicating genotype phenotype associations. *Nature* 447:655–660.
2. Ioannidis JPA (2001) Why most published research findings are false. *PLoS Med* 2:e124.
3. Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG (2001) Replication validity of genetic association studies. *Nat Genet* 29:306–309.
4. Donoho D, Jin J (2008) Higher Criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proc Natl Acad Sci USA* 105:14790–14795.
5. Donoho D, Jin J (2008) Feature selection by Higher Criticism thresholding: Optimal phase diagram. arXiv:0812.2263v1 [math.ST].
6. Hall P, Jin J (2009) Innovated Higher Criticism for detecting sparse signals in correlated noise. arXiv:0902.3837v1 [math.ST].
7. Ingster YI (1997) Some problems of hypothesis testing leading to infinitely divisible distribution. *Math Methods Stat* 6:47–69.
8. Donoho D, Jin J (2004) Higher criticism for detecting sparse heterogeneous mixtures. *Ann Stat* 32:962–994.
9. Jin J (2003) Detecting and estimating sparse mixtures. PhD thesis (Department of Statistics, Stanford Univ, Stanford).
10. Fan J, Fan Y (2008) High-dimensional classification using features annealed independence rules. *Ann Stat* 36:2605–2637.
11. Le Cam L, Yang G (2000) *Asymptotics in Statistics* (Springer, New York).
12. Fan J (1996) Testing of significance based on wavelet thresholding and Neyman's truncation. *J Am Stat Assoc* 91:674–688.
13. Fan J, Song R (2009) Sure independence screening in generalized linear models with NP-dimensionality. arXiv:0903.5255 [math.ST].
14. Hall P, Pittelkow Y, Ghosh M (2008) Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes. *J R Stat Soc B* 70:158–173.
15. Ingster Y, Pouet C, Tsybakov AB (2009) Sparse classification boundaries. arXiv: 0903.4807 [math.ST].