Feature selection by higher criticism thresholding achieves the optimal phase diagram

BY DAVID DONOHO1 AND JIASHUN JIN2,*

1Department of Statistics, Stanford University, Stanford, CA 94305, USA
2Statistics Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

We consider two-class linear classification in a high-dimensional, small-sample-size setting. Only a small fraction of the features are useful, these being unknown to us, and each useful feature contributes weakly to the classification decision. This was called the rare/weak (RW) model in our previous study (Donoho, D. & Jin, J. 2008 Proc. Natl Acad. Sci. USA 105, 14790–14795). We select features by thresholding feature Z-scores. The threshold is set by higher criticism (HC). For 1 ≤ i ≤ N, let π_i denote the p-value associated with the i-th Z-score and π(1) denote the 1-th order statistic of the collection of p-values. The HC threshold (HCT) is the order statistic of the Z-score corresponding to index i maximizing (i/N − π(i))/(i/N(1−i/N)). The ideal threshold optimizes the classification error. In that previous study, we showed that HCT was numerically close to the ideal threshold.

We formalize an asymptotic framework for studying the RW model, considering a sequence of problems with increasingly many features and relatively fewer observations. We show that, along this sequence, the limiting performance of ideal HCT is essentially just as good as the limiting performance of ideal thresholding. Our results describe two-dimensional phase space, a two-dimensional diagram with coordinates quantifying ‘rare’ and ‘weak’ in the RW model. The phase space can be partitioned into two regions—one where ideal threshold classification is successful, and one where the features are so weak and so rare that it must fail. Surprisingly, the regions where ideal HCT succeeds and fails make exactly the same partition of the phase diagram. Other threshold methods, such as false (feature) discovery rate (FDR) threshold selection, are successful in a substantially smaller region of the phase space than either HCT or ideal thresholding. The FDR and local FDR of the ideal and HC threshold selectors have surprising phase diagrams, which are also described. Results showing the asymptotic equivalence of HCT with ideal HCT can be found in a forthcoming paper (Donoho, D. & Jin, J. In preparation).

Keywords: asymptotic rare/weak model; false discovery rate; linear classification; phase diagram; Fisher’s separation measure; feature selection by thresholding

*Author for correspondence (jiashun@stat.cmu.edu).


One contribution of 11 to a Theme Issue ‘Statistical challenges of high-dimensional data’.
1. Introduction

The modern era of high-throughput data collection creates data in abundance. Some devices—spectrometers and gene chips come to mind—automatically generate measurements on thousands of standard features of each observational unit.

What has not changed in science is the difficulty of obtaining good observational units. High-throughput devices do not help us to find and enrol qualified patients, or to catch exotic butterflies or to observe primate mating behaviour. So the number of observational units—be they patients, butterflies or matings—has not increased; it stays in the dozens or hundreds. However, each of those few observational units can now be subjected to a large battery of automatic feature measurements.

Many of those automatically measured features will have little relevance to any given project. This new era of feature glut poses a needle-in-a-haystack problem: we must detect a relatively few valuable features among many useless ones. Unfortunately, the combination of small sample sizes (few observational units) and high dimensions (many feature measurements) makes it hard to tell needles from straw.

Orthodox statistical methods assumed quite a different set of conditions: more observations than features, and all features highly relevant. Modern statistical research has started to address this modern, unorthodox setting; such research comprised much of the activity in the recent six-month Newton Institute programme Statistical Theory and Methods for Complex, High-Dimensional Data.

This paper examines the challenges that feature glut poses to linear classification schemes. A new feature selection principle—higher criticism (HC) thresholding—and a new optimality property—optimality of the phase diagram—will be introduced.

(a) Multivariate normal classification

In the standard model of linear classifier training, we have labelled training samples \((Y_i, X_i), i = 1, \ldots, n\), where each label \(Y_i\) is \(\pm 1\) and each feature vector \(X_i \in \mathbb{R}^p\). For simplicity, let us assume that the training set contains equal numbers of +1 and −1 and that the feature vectors \(X_i \in \mathbb{R}^p\) obey \(X_i \sim N(Y_i \mu, \Sigma)\), \(i = 1, \ldots, n\), for an unknown mean contrast vector \(\mu \in \mathbb{R}^p\); here \(\Sigma\) denotes the feature covariance matrix and \(n\) is the training set size. In this simple setting, one normally uses linear classifiers to classify an unlabelled test vector \(X\), taking the general form \(L(X) = \sum_{j=1}^{p} w(j) X(j)\), with ‘feature weights’ \(w = (w(j) : j = 1, \ldots, p)\).

Classical theory going back to R. A. Fisher shows that the optimal classifier has feature weights \(w \propto \Sigma^{-1} \mu\) (Anderson 2003). However, modern feature glut renders Fisher’s formula inapplicable.

(b) When \(p\) is larger than \(n\)

Frequently, in modern settings, the number \(p\) of measurements exceeds the number \(n\) of observational units. Naive application of the formula \(w \propto \Sigma^{-1} \mu\) to empirical data in the \(p > n\) setting is then problematic, at a minimum, because
the matrix of empirical feature covariances $\hat{\text{Cov}}_{n,p}(X)$ is not invertible. Even the generalized inverse $\hat{\text{Cov}}_{n,p}(X)^\dagger$ gives weights $\hat{w}_{\text{naive}} \propto \hat{\text{Cov}}_{n,p}(X)^\dagger \hat{\text{Cov}}_{n,p}(Y, X)$, yielding ‘noisy’ classifiers with low accuracy.

A by-now standard response to this problem is simply to ignore feature covariances, and to standardize the features to mean 0 and variance 1. In effect, one pretends that the feature covariance matrix $\Sigma$ is the identity matrix, and uses the formula $w(j) \propto \text{Cov}(Y, X(j))$ (Bickel & Levina 2004; Fan & Fan 2008). Even after this reduction, further challenges remain.

(c) When features are rare and weak

When feature glut arises from promiscuous measurement, researchers cannot say in advance which features will be useful in a given project. Moreover, although one hopes to find a single ‘silver bullet’ feature that reliably classifies, in all too many projects the useful features are relatively rare and individually quite weak.

Such thinking motivated the following rare/weak (RW) feature model in Donoho & Jin (2008). Under this model, the following hold.

(i) Useful features are rare. The contrast vector $\mu$ is non-zero in only $k$ out of $p$ elements, where $\epsilon = k/p$ is small, i.e. close to zero. As an example, think of $p = 10000$, $k = 100$, so $\epsilon = k/p = 0.01$.

(ii) Useful features are weak. The non-zero elements of $\mu$ have common amplitude $\mu_0$, which is assumed not to be ‘large’. ‘Large’ can be measured using $\tau = \sqrt{n\mu_0}$; values of $\tau$ in the range of 2–4 imply that corresponding values of $\mu$ are not large.

Since the elements $X(j)$ of the feature vector where the class contrast $\mu(j) = 0$ are entirely uninformative about the value of $Y(j)$, only the $k$ features where $\mu(j) = \mu_0$ are useful. The problem is how to identify and benefit from those rare/weak features.

We speak of $\epsilon$ and $\tau$ as the sparsity and strength parameters for the RW model, and refer to the RW($\epsilon$, $\tau$) model. Models with a ‘sparsity’ parameter $\epsilon$ are common in estimation settings (Donoho et al. 1992; Donoho & Johnstone 1994; Johnstone & Silverman 2004), but not with the feature strength constraint $\tau$. Also closely related to the RW model is work in multiple testing by Ingster and other authors (Ingster 1997; Jin 2003; Donoho & Jin 2004).

(d) Feature selection by thresholding

Feature selection—i.e. working only with an empirically selected subset of features—is a standard response to feature glut. We are supposing, as announced in §1b, that feature correlations can be ignored and that features are already standardized to variance 1. We therefore focus on the vector of feature $Z$-scores, with components $Z(j) = n^{-1/2} \sum_i Y_i X_i(j)$, $j = 1, \ldots, p$. These are the $Z$-scores of two-sided tests of $H_{0,j}$: $\text{Cov}(Y, X(j)) = 0$. Under our assumptions, $Z \sim N(\theta, I_p)$ with $\theta = \sqrt{n\mu}$ and $\mu$ the feature contrast vector. Features with non-zero $\mu(j)$ typically have significantly non-zero $Z(j)$, while all other features will have $Z(j)$ values largely consistent with the null hypothesis $\mu(j) = 0$. In such a setting, selecting features with $Z$-scores above a threshold makes sense.
We identify three useful threshold functions: \( \eta^*_t(Z) \), with \( \star \in \{ \text{clip, hard, soft} \} \). These are: clipping, \( \eta^\text{clip}_t(Z) = \text{sgn}(Z) \cdot 1_{\{|Z| > t\}} \), which ignores the size of the \( Z \)-score, provided it is large; hard thresholding, \( \eta^\text{hard}_t(Z) = Z \cdot 1_{\{|Z| > t\}} \), which uses the size of the \( Z \)-score, provided it is large; and soft thresholding, \( \eta^\text{soft}_t(z) = \text{sgn}(Z)(|Z| - t)_+ \), which uses a shrunken \( Z \)-score, provided it is large.

**Definition 1.1.** Let \( \star \in \{ \text{soft, hard, clip} \} \). The threshold feature selection classifier makes its decision based on \( L^*_t <> 0 \), where \( \hat{L}^*_t(X) = \sum_{j=1}^p \hat{w}^*_t(j)X(j) \), and \( \hat{w}^*_t(j) = \eta^*_t(Z(j)) \), \( j = 1, \ldots, p \).

In other words, the classifier sums across features with large training set \( Z \)-scores, and a simple function of the \( Z \)-score generates the feature weight.

Several methods for linear classification in bioinformatics follow this approach. The shrunken centroids method (Tibshirani et al. 2002) is a variant of soft thresholding in this two-class setting. The highly cited methods in Golub et al. (1999) and Hedenfalk et al. (2001) are variants of hard thresholding. Clipping makes sense in the theoretical setting of the RW model (since then the useful features all have the same strength) and is simpler to analyse than the other nonlinearities.

One crucial question remains: how to choose the threshold based on the data? Popular methods for threshold choice include cross-validation (Tibshirani et al. 2002), control of the false discovery rate (Abramovich & Benjamini 1995; Benjamini & Hochberg 1995; Abramovich et al. 2006; Donoho & Jin 2006) and control of the local false discovery rate (Efron et al. 2001).

**Higher criticism**

Recent work on multiple comparisons offers a new method of threshold choice.

(i) Higher criticism testing

Suppose that we have a collection of \( N \) \( p \)-values \( \pi_i \), which, under the global null hypothesis, are uniformly (independent and identically) distributed: \( \pi_i \sim \text{i.i.d.} U[0, 1] \). Consider the order statistics \( \pi_{(1)} \leq \pi_{(2)} \leq \cdots \leq \pi_{(N)} \). Under the null hypothesis, these order statistics have the usual properties of uniform order statistics, including the asymptotic normality \( \pi_{(i)} \sim \text{approx Normal}(i/N, i/N(1 - i/N)) \). HC forms a \( Z \)-score comparing \( \pi_{(i)} \) with its mean under the null, and then maximizes over a wide range of \( i \).

**Definition 1.2 (HC testing (Donoho & Jin 2004)).** The higher criticism objective is

\[
\text{HC}(i; \pi_{(i)}) = \sqrt{N} \frac{i/N - \pi_{(i)}}{\sqrt{i/N(1 - i/N)}}. \tag{1.1}
\]

Fix \( \alpha_0 \in (0, 1) \) (e.g. \( \alpha_0 = 0.10 \)). The higher criticism test statistic is \( \text{HC}^* = \max_{1 \leq i \leq N} \text{HC}(i; \pi_{(i)}) \).

HC seems insensitive to the selection of \( \alpha \), in rare/weak situations; here we always use \( \alpha_0 = 0.10 \).

Expressed in words, we look for the largest standardized discrepancy for any \( \pi_{(i)} \) between the observed behaviour and the expected behaviour under the null. When this is large, the whole collection of \( p \)-values is not consistent with the
Figure 1. Illustration of HC thresholding. (a) The ordered |Z|-scores. (b) The corresponding ordered p-values in a PP plot. (c) The HC objective function in equation (1.1); this is largest at $\hat{i} \approx 0.007p$. The x-axes are $i/p$. Vertical (red) lines indicate $\pi(\hat{i})$ in (b), and $|Z|_{(\hat{i})}$ in (a).

global null hypothesis. The phrase ‘higher criticism’ is due to John Tukey, and reflects the shift in emphasis from single test results to the whole collection of tests.

(ii) Higher criticism thresholding

We return to the classification setting in the previous sections. We have a vector of feature Z-scores $(Z(j), j = 1, \ldots, p)$. To apply HC notions, translate Z-scores into two-sided p-values, and maximize the HC objective over index $i$ in the appropriate range. Define the feature p-values $\pi_i = \text{Prob}\{|N(0,1)| > |Z(i)|\}, i = 1, \ldots, p$; and define the increasing rearrangement $\pi_{(i)}$, the HC objective function $HC(i; \pi_{(i)})$, and the increasing rearrangement $|Z|_{(i)}$ correspondingly. Here is our proposal.

**Definition 1.3 (HC thresholding).** Apply the HC procedure to the feature p-values. Let the maximum HC objective be achieved at index $\hat{i}$. The higher criticism threshold (HCT) is the value $\hat{\tau}_{HC} = |Z|_{(\hat{i})}$. The HC threshold feature selector selects features with Z-scores exceeding $\hat{\tau}_{HC}$ in magnitude.

Figure 1 illustrates the procedure. Figure 1a shows a sample of Z-scores, figure 1b shows a probability–probability (PP) plot of the corresponding ordered p-values versus $i/p$, and figure 1c shows a standardized PP plot. The standardized PP plot has its largest deviation from 0 at $\hat{i}$; and this generates the threshold value.
(iii) Previously reported results for higher criticism threshold

Donoho & Jin (2008) reported several findings on the behaviour of HCT based on numerical and empirical evidence. In the RW model, define the ideal threshold as the optimal threshold based on full knowledge of the RW parameters $\epsilon$ and $\mu$ (see §2 below).

(i) The HCT threshold is numerically very close to the ideal threshold.
(ii) In the case of very weak feature $Z$-scores, HCT has a false feature discovery rate (FDR) substantially higher than other popular approaches, but a feature missed detection rate (MDR) substantially lower than those other approaches.
(iii) FDR and MDR of HCT closely match those of the ideal threshold.

In short, HCT has operating characteristics unlike those of popular thresholding schemes such as FDR thresholding (Abramovich & Benjamini 1995; Abramovich et al. 2006) and Bonferroni thresholding, but very similar to the ideal ones.

(f) Asymptotic rare/weak model, and the phase diagram

This paper reports a large-sample analysis of HCT and the RW model. The number of observations $n$ and the number of features $p$ will tend to infinity in a linked fashion, with $n$ remaining very small compared to $p$. (Empirical results in Donoho & Jin (2008) show that large-$p$ theory is applicable at moderate $n$ and $p$).

Definition 1.4. The phrase asymptotic rare/weak (ARW) model refers to the following combined assumptions.

(i) Asymptotic setting. We consider a sequence of problems, with the number of observations $n$ and the number of features $p$ both tending to $\infty$.
(ii) $p$ dramatically larger than $n$. Along this sequence, $n \sim c(\log p)^\gamma$ for some positive constants $c$ and $\gamma$, so there are dramatically more features per observational unit than there are observational units.
(iii) Increasing rarity. The sparsity $\epsilon$ tends to 0: $\epsilon = p^{-\beta}, 0 < \beta < 1$.
(iv) Decreasing strength. The strength $\tau$ varies with $n$ and $p$ according to $\tau = \sqrt{2r \log p}, 0 < r < 1$.

The symbol ARW$(r, \beta; c, \gamma)$ refers to the model combining these assumptions.

In this model, because $r < 1$, useful features are individually too weak to detect and, because $0 < \beta < 1$, useful features are increasingly rare with increasing $p$, while increasing in total number with $p$. It turns out that $c$ and $\gamma$ are incidental, while $r$ and $\beta$ are the driving parameters. Hence, we always simply write ARW$(r, \beta)$ below.

For many choices of $(r, \beta)$, asymptotically successful classification is possible; for another large family of choices, it is impossible. To understand this fact, we use the concept of phase space, the two-dimensional domain, $0 < r, \beta < 1$. We show that this domain is partitioned into two regions or ‘phases’. In the ‘impossible’ phase, useful features are so rare and so weak that classification is
Figure 2. Phase diagram. The curve \( r = \rho^*(\beta) \) divides phase space into failure and success regions, and the latter further splits into three different regions I, II and III. Numbers in the brackets show limits of \( \tilde{FDR} \) and \( \tilde{Lfd} \) at the ideal HCT as in theorem 3.7.

Asymptotically impossible even with the ideal choice of threshold. In the ‘possible’ phase, successfully separating the two groups is indeed possible—if one has access to the ideal threshold. Figure 2 displays this domain and its partition into phases. Because of the partition into two phases, we also call this display the phase diagram. An explicit formula for the graph \( r = \rho^*(\beta) \) bounding these phases is given in equation (3.2) below.

The phase diagram provides a convenient platform for comparing different procedures. A threshold choice is optimal if it gives the same partition of phase space as the one obtained with the ideal choice of threshold.

How does HCT compare to the ideal threshold, and what partition in the phase space does HCT yield? For reasons of space, we focus in this paper on the ideal HC threshold, which is obtained upon replacing the empirical distribution of feature \( Z \)-scores by its expected value. The ideal HC threshold is thus the one that HCT is ‘trying to estimate’.

The central surprise of our story is that ideal HCT behaves surprisingly well: the partition of phase space describing the two regions where ideal thresholding fails and/or succeeds also describes the two regions where HCT fails and/or succeeds. The situation is depicted in table 1: here ‘succeeds’ means asymptotically zero misclassification rate, and ‘fails’ means asymptotically 50 per cent misclassification rate.

In this sense of size of regions of success, HCT is just as good as the ideal threshold. Such statements cannot be made for some other popular thresholding schemes, such as false discovery threshold selection. Even the very popular
Table 1. Comparison of ideal threshold and ideal HCT.

<table>
<thead>
<tr>
<th>region</th>
<th>property of ideal threshold</th>
<th>property of ideal HCT</th>
</tr>
</thead>
<tbody>
<tr>
<td>$r &lt; \rho^*(\beta)$</td>
<td>ideal threshold classifier fails</td>
<td>ideal HCT fails</td>
</tr>
<tr>
<td>$r &gt; \rho^*(\beta)$</td>
<td>ideal threshold classifier succeeds</td>
<td>ideal HCT succeeds</td>
</tr>
</tbody>
</table>

cross-validated choice of threshold will fail if the training set size is bounded, while HCT will still succeed in the RW model in that case.

\((g)\) Contents

The paper is organized as follows. Section 2 introduces a functional framework and several ideal quantities. These include the proxy classification error where Fisher’s separation (Sep) plays a key role, the ideal threshold as a proxy for the optimal threshold, and the ideal HCT as a proxy of the HCT. Section 3 introduces the main results on the asymptotic behaviour of the HC threshold under the asymptotic RW model, and the focal point is the phase diagram. Section 4 outlines the basic idea behind the main results followed by the proofs. Section 5 discusses the connection between the ideal threshold and the ideal HCT. Section 6 discusses the ideal behaviour of Bonferroni threshold feature selection and FDR-controlling feature selection. Section 7 discusses the link between ideal HCT and ordinary HCT, the finite-p phase diagram, and other appearances of HC in recent literature.

This paper has three companions: Donoho & Jin (2008, in preparation) and Jin (2009). The full proof of a broader set of claims, with a considerably more general treatment, will appear in Donoho & Jin (in preparation) (see §7).

2. Fisher’s separation functional and ideal threshold

Suppose $L$ is a fixed, non-random linear classifier, with decision boundary $L <> 0$. Will $L$ correctly classify the future realization $(Y, X)$? Suppose that $Y = \pm 1$ are equiprobable and $X \sim N(Y \mu, I_p)$. The misclassification probability can be written as:

$$P\{YL(X) < 0 \mid \mu\} = \Phi \left( -\frac{1}{2} \text{Sep}(L; \mu) \right), \quad (2.1)$$

where $\Phi$ denotes the standard normal distribution function and Sep($L; \mu$) measures the standardized interclass distance:

$$\text{Sep}(L; \mu) = \frac{E[L(X) \mid Y = 1] - E[L(X) \mid Y = -1]}{\text{SD}(L(X))} = \frac{2\langle w, \mu \rangle}{\langle w \rangle^2} = \frac{2\langle w, \mu \rangle}{\|w\|^2}. \quad (2.2)$$

The ideal linear classifier $L_\mu$ with feature weights $w \propto \mu$ and decision threshold $L_\mu <> 0$ implements the likelihood ratio test. It also maximizes Sep, since, for every other linear classifier $L$, $\text{Sep}(L; \mu) \leq \text{Sep}(L_\mu; \mu) = 2\|\mu\|^2$. 

*Phil. Trans. R. Soc. A* (2009)
Threshold selection rules give random linear classifiers: the classifier weight vector \(w\) is a random variable, because it depends on the \(Z\)-scores of the realized training sample. If \(L_Z\) denotes a linear classifier constructed based on such a realized vector of \(Z\)-scores, then the misclassification error can be written as

\[
\text{Err}(L_Z, \mu) = \Phi\left(-\frac{1}{2} \text{Sep}(L_Z; \mu)\right);
\]

this is a random variable depending on \(Z\) and on \(\mu\). Heuristically, because there is a large number of coordinates, some statistical regularity appears, and we anticipate that random quantities can be replaced by expected values. We proceed as if

\[
\text{Sep}(L_Z; \mu) \approx \frac{2E_{Z|\mu} \langle w, \mu \rangle}{(E_{Z|\mu} \|w\|^2)^{1/2}},
\]

where the expectation is over the conditional distribution of \(Z\) conditioned on \(\mu\). Next, let \(I_1, I_2, \ldots, I_p\) be independent Bernoulli random variables with parameter \(\epsilon\), and \(\mu_j = I_j \mu_0\), so that \(\mu\) has about \(\epsilon p\) non-zero coordinates, in random locations. Since \(w_i = \eta_i(Z_i)\), we write heuristically

\[
E_{Z|\mu} \langle w, \mu \rangle \approx p\epsilon \mu_0 E\{\eta_1(Z_1) | \mu_1 = \mu_0\},
\]

\[
E_{Z|\mu} \langle w, w \rangle \approx p(\epsilon E\{\eta_1^2(Z_1) | \mu_1 = \mu_0\} + (1 - \epsilon) E\{\eta_1^2(W) | \mu_1 = 0\}).
\]

Here \(Z_1\) and \(\mu_1\) are the first coordinates of \(Z\) and \(\mu\), respectively, and \(\mu_0\) is a constant as before.

**Definition 2.1.** Let the threshold be fixed and chosen independently of the training set. In the RW(\(\epsilon, \tau\)) model, we use the following expression for proxy separation:

\[
\widetilde{\text{Sep}}(t; \epsilon, \tau) = 2A/\sqrt{B},
\]

where

\[
A(t, \epsilon, \tau) = \epsilon \tau E\{\eta_1(\tau + W)\},
\]

\[
B(t, \epsilon, \tau) = \epsilon E\{\eta_1^2(\tau + W)\} + (1 - \epsilon) E\{\eta_1^2(W)\},
\]

and \(W\) denotes a standard normal random variable. By proxy classification error we mean

\[
\widetilde{\text{Err}}(t; \epsilon, \tau, p, n) = \Phi\left(-\frac{1}{2} \sqrt{\frac{p}{n}} \widetilde{\text{Sep}}(t, \epsilon, \tau)\right).
\]

Normalizations are chosen so that, in large samples,

\[
\text{Sep}(L_Z, \mu) \approx \sqrt{\frac{p}{n}} \widetilde{\text{Sep}}(t, \epsilon, \tau).
\]

While, normally, we expect averages to ‘behave like expectations’ in large samples, the word proxy reminds us that there is a difference (presumably small). Software to compute proxy expressions has been developed by the authors, and some numerical results were reported in Donoho & Jin (2008).
Of course, the rationale for our interest in these proxy expressions is our heuristic understanding that they accurately describe the exact large-sample behaviour of certain threshold selection schemes. This issue is settled in the affirmative, after considerable effort, in Donoho & Jin (in preparation).

(b) Certainty equivalent threshold functionals

In general, the best threshold to use in a given instance of the RW model depends on both $\epsilon$ and $\tau$. It also depends on the specific realization of $\mu$ and even of $Z$. However, the dependence on $\mu$ and $Z$ is simply ‘noise’ that goes away in large samples, while the dependence on $\epsilon$ and $\tau$ remains.

Definition 2.2. The ideal threshold functional $T_{\text{ideal}}(\epsilon, \tau)$ maximizes the proxy separation $T_{\text{ideal}}(\epsilon, \tau) = \text{arg max}_t \hat{\text{Sep}}(t, \epsilon, \tau)$.

Heuristically, $T_{\text{ideal}}$ represents a near-optimal threshold in all sufficiently large samples; it is what we ‘ought’ to be attempting to use.

Definition 2.3 (Folding). The following concepts and notation will be used in connection with distributions of absolute values of random variables.

(i) The half-normal distribution function $\Psi(t) = P\{|N(0,1)| \leq t\}$.
(ii) The non-central half-normal distribution $\Psi_\tau(t) = P\{|N(\tau,1)| \leq t\}$.
(iii) Given a distribution function $F$, the folded distribution is given by $G(t) = F(t) - F(-t)$. The half-normal is the folded version of the standard normal, and the non-central half-normal is the folded version of a normal with unit standard deviation and non-zero mean equal to the non-centrality parameter.
(iv) Let $F_{\epsilon,\tau}$ denote the two-point mixture $F_{\epsilon,\tau}(t) = (1 - \epsilon)\Phi(t) + \epsilon\Phi(t - \tau)$; then $G_{\epsilon,\tau}$ denotes the corresponding folded distribution given by $G_{\epsilon,\tau}(t) = (1 - \epsilon)\Psi_0(t) + \epsilon\Psi_\tau(t)$.

We now define an HCT functional representing the target that HC thresholding aims for. For a distribution function $F$, the survival function $\tilde{F} = 1 - F$ is its complement.

Definition 2.4. Let $F$ be a distribution function that is not the standard normal $\Phi$. At such a distribution, we define the HCT functional by

$$T_{\text{HC}}(F) = \text{arg max}_{t > t_0} (\tilde{G}(t) - \tilde{\Psi}(t))/\sqrt{\tilde{G}(t) \cdot \tilde{G}(t)}.$$ 

Here $G$ is the folding of $F$, and $t_0$ is a fixed parameter of the HCT method (e.g. $t_0 = \Phi^{-1}(0.1)$). The HC threshold in the RW$(\epsilon, \tau)$ model may be written, in an abuse of notation, as $T_{\text{HC}}(\epsilon, \tau)$, meaning $T_{\text{HC}}(F_{\epsilon,\tau})$.

Let $F_{n,p}$ denote the usual empirical distribution of the feature $Z$-scores $Z_i$. The HCT of definition 1.3 can be written as $T_{\text{HC}}^{n,p} = T_{\text{HC}}(F_{n,p})$. Let $F$ denote the expected value of $F_{n,p}$; then $T_{\text{HC}}(F)$ will be called the ideal HC threshold. Heuristically, we expect the usual sampling fluctuations and that $T_{\text{HC}}(F_{n,p}) \approx T_{\text{HC}}(F)$ with a discrepancy decaying as $p$ and $n$ increase. This issue is carefully
Feature selection by HC thresholding

considered in the companion paper (Donoho & Jin in preparation), which shows that the empirical HC threshold in the ARW model indeed closely matches the ideal HC threshold.

For comparison purposes, we considered two other threshold schemes. First, (ideal) false discovery rate thresholding (FDRT). For a threshold $t$, and parameters $(p, \epsilon, \tau)$, the expected number of useful features selected is

$$E(\text{TP})(t; \epsilon, \tau, p) = p \epsilon \tilde{G}_{\epsilon, \tau}(t),$$

and the expected number of useless features selected is

$$E(\text{FP})(t; \epsilon, \tau, p) = p(1 - \epsilon) \tilde{\Psi}(t).$$

Here, TP/FP stand for true/false positives. Let $\text{TPR}(t) = p^{-1}E(\text{TP})(t)$ denote the expected rate of useful features above threshold and $\text{FPR}(t) = p^{-1}E(\text{FP})(t)$ the expected rate of useless features above threshold. By analogy with our earlier heuristic, we define the proxy false discovery rate

$$\tilde{\text{FDR}}(t; \epsilon, \tau, p) = \frac{\text{FPR}(t)}{\text{TPR}(t) + \text{FPR}(t)}.$$

The term ‘proxy’ reminds us that

$$E(\text{FP})(t)/E(\text{TP})(t) + E(\text{FP})(t)) \neq E[\text{FP}(t)/\text{TP}(t) + \text{FP}(t))],$$

although for large $p$ the difference will often be small.

We define the FDRT-$\alpha$ functional by

$$T_{\text{FDR}, \alpha}(\epsilon, \tau) = \min\{t : \tilde{\text{FDR}}(t; \epsilon, \tau, p) < \alpha, t > t_0\}.$$

Heuristically, this is the threshold that FDRT is ‘trying’ to learn from noisy empirical data. We will also need the proxy local FDR:

$$\tilde{\text{Lfdr}}(t; \epsilon, \tau, p) = \text{FPR}'(t)/(\text{TPR}'(t) + \text{FPR}'(t)).$$

Here $\text{FPR}'$ denotes the derivative of FPR, which exists, using smoothness properties of $\Psi_0$; similarly for $\text{TPR}'$ and $\Psi_{\tau}$. Intuitively, $\tilde{\text{Lfdr}}(t)$ denotes the expected fraction of useless features among features having observed $Z$-scores near level $t$.

Secondly, we considered Bonferroni-based thresholding: $T_{\text{Bon}} = \phi^{-1}(p^{-1})$. This threshold level is set at the level that would cause on average one false alarm in a set of $p$ null cases. Figs 2 and 3 of Donoho & Jin (2008) presented numerical calculations of these functionals and their separation behaviour in two cases: (a) $p = 10^4$, $\epsilon = 0.01$, and (b) $p = 10^6$, $\epsilon = 0.0001$. Although our calculations are exact numerical finite-$p$ calculations, we remark that they correspond to sparsity exponents $\beta = 1/2$ and $\beta = 2/3$, respectively. Those figures show the following.

(i) There is a very close numerical approximation of the HCT to the ideal threshold, not just at large $\tau$ but also even at quite small $2 < \tau < 3$.

(ii) FDR and Bonferroni thresholding behave differently from ideal thresholding and ideal HCT.

(iii) The separation behaviour of the HCT is nearly ideal. For the constant FDR rules, the separation behaviour is close to ideal at some $\tau$ but becomes noticeably sub-ideal at other $\tau$. 

Phil. Trans. R. Soc. A (2009)
(iv) The false discovery rate behaviour of HCT and ideal thresholding depends on $\tau$. At small $\tau$, both rules tolerate a high FDR, while, at large $\tau$, both rules obtain a small FDR.

(v) The missed detection rate of HCT and ideal thresholding also depends on $\tau$. At small $\tau$, the missed detection rate is high, but noticeably less than 100 per cent. At large $\tau$, the missed detection rate falls, but remains noticeably above 0 per cent. In contrast, the MDR for FDR procedures is essentially 100 per cent for small $\tau$ and falls below that of HCT/ideal for large $\tau$.

These numerical examples illustrate the idealized behaviour of different procedures. We can think of the HCT functional as the threshold that is being estimated by the actual HCT rule. On an actual dataset sampled from the underlying $F$, the HC threshold will behave differently, primarily due to stochastic fluctuations $F_{n,p} \approx F$. Nevertheless, the close approximation of the HC threshold to the ideal one is striking and, to us, compelling.

3. Behaviour of ideal threshold, asymptotic rare/weak model

We now study the ideal threshold in the asymptotic RW model of definition 1.4. That is, we fix parameters $r, \beta$ in that model and study the choice of threshold $t$ maximizing class separation.

We first make precise a structural fact about the ideal threshold, first observed informally in Donoho & Jin (2008).

Definition 3.1 (ROC curve). The feature detection receiver operating characteristic curve (ROC) is the curve parametrized by $(\text{FPR}(t), \text{TPR}(t))$. The tangent to this curve at $t$ is $\tan(t) = \text{TPR}'(t)/\text{FPR}'(t)$, and the secant is $\sec(t) = \text{TPR}(t)/\text{FPR}(t)$. Note that, in the RW($\epsilon, \tau$) model, TPR, FPR, tan and sec all depend on $t, \epsilon, \tau$ and $p$, although we may, as here, indicate only a dependence on $t$.

Theorem 3.2 (Tangent–secant rule). In the ARW($r, \beta$) model definition 1.4, we have

$$\frac{\tan(T_{\text{ideal}})}{1 + \tan(T_{\text{ideal}})} \sim \frac{1}{2}\left(1 + \frac{\sec(T_{\text{ideal}})}{1 + \sec(T_{\text{ideal}})}\right), \quad p \longrightarrow \infty. \quad (3.1)$$

Definition 3.3 (Success region). The region of asymptotically successful ideal threshold feature selection in the ($\beta, r$) plane is the interior of the subset where, under the ARW($r, \beta$) model of definition 1.4, the ideal threshold choice $T_{\text{ideal}}(\epsilon, \tau)$ obeys $\tilde{\text{Err}}(T_{\text{ideal}}(\epsilon, \tau); \epsilon, \tau, p) \rightarrow 0, \quad p \rightarrow \infty$.

The interesting range involves $(\beta, r) \in [0, 1]^2$. The following function is important for our analysis, and has previously appeared in important roles in other (seemingly unrelated) problems (see §7):

$$\rho^*(\beta) = \begin{cases} 0, & 0 < \beta \leq 1/2, \\ \beta - 1/2, & 1/2 < \beta \leq 3/4, \\ (1 - \sqrt{1 - \beta})^2, & 3/4 < \beta < 1. \end{cases} \quad (3.2)$$
As it turns out, it marks the boundary between success and failure for threshold feature selection. (The correct interpretation of $\rho^*(\beta) = 0$ for $\beta \in (0, 1/2]$ is simply that the critical value of $\tau$ separating success from failure is $o(\sqrt{2\log p})$.)

**Theorem 3.4 (Existence of phases).** The success region is precisely $r > \rho^*(\beta)$, $0 < \beta < 1$. In the interior of the complementary region, $r < \rho^*(\beta)$, $1/2 < \beta < 1$, even the ideal threshold cannot send the proxy separation to infinity with increasing $(n, p)$.

**Definition 3.5 (Regions I, II, III).** The success region can be split into three regions, referred to here and below as regions I–III. The interiors of the regions are as follows (figure 2):

I. $0 < r \leq \beta/3$ and $1/2 < \beta < 3/4$; $r > \rho^*(\beta)$.
II. $\beta/3 < r \leq \beta$ and $1/2 < \beta < 1$; $r > \rho^*(\beta)$.
III. $\beta < r < 1$ and $1/2 < \beta < 1$; $r > \rho^*(\beta)$.

In the asymptotic RW model, the optimal threshold must behave asymptotically like $\sqrt{2q\log p}$ for a certain $q = q(r, \beta)$. Surprisingly, we need not have $q = r$.

**Theorem 3.6 (Formula for ideal threshold).** Under the asymptotic RW model $ARW(r, \beta)$, with $r > \rho^*(\beta)$, the ideal threshold has the form $T_{\text{ideal}}(\epsilon, \tau) \sim \sqrt{2q^*\log p}$ where

$$q^* = \begin{cases} 4r, & \text{region I,} \\ (\beta + r)/4r, & \text{regions II, III.} \end{cases} \quad (3.3)$$

Note in particular that, in regions I and II, $q^* > r$, and hence $T_{\text{ideal}}(\epsilon, \tau) > \tau$. Although the features truly have strength $\tau$, the threshold is best set higher than $\tau$. An apparently similar result was found in the detection setting (Donoho & Jin 2004), but the formula for $q^*$ was subtly different from equation (3.3).

We now turn to FDR properties. The tangent–secant rule implies immediately

$$\frac{\widehat{LF}_{\text{d}}(T_{\text{ideal}})}{(1 + \text{FDR}(T_{\text{ideal}}))/2} \longrightarrow 1, \quad p \longrightarrow \infty. \quad (3.4)$$

Hence, any result about FDR is tied to one about local FDR, and vice versa.

**Theorem 3.7.** Under the asymptotic RW model $ARW(r, \beta)$, at the ideal threshold $T_{\text{ideal}}(\epsilon, \tau)$, proxy FDR and proxy local FDR obey

$$\widehat{\text{FDR}}(T_{\text{ideal}}, \epsilon, \tau) \longrightarrow b(\beta, r), \quad \widehat{LF}_{\text{d}}(T_{\text{ideal}}, \epsilon, \tau) \longrightarrow \frac{1 + b(\beta, r)}{2}, \quad p \longrightarrow \infty, \quad (3.5)$$

where

$$b(\beta, r) = \begin{cases} 1, & \text{region I,} \\ (r + \beta)/(4r), & \text{region II} \quad \text{(note } \beta/3 < r < \beta), \\ 1/2, & \text{region III.} \end{cases}$$

Several aspects of the above solution are of interest.
(i) Threshold elevation. The threshold $\sqrt{2q^*\log p}$ is significantly higher than $\sqrt{2r\log p}$ in regions I and II. Instead of looking for features at the amplitude they can be expected to have, we look for them at much higher amplitudes.

(ii) Fractional harvesting. Outside region III, we are selecting only a small fraction of the truly useful features.

(iii) False discovery rate. Outside region III, we actually have a very large false discovery rate, which is very close to 1 in region I. Surprisingly, even though most of the selected features are useless, we still correctly classify!

(iv) Training versus test performance. The quantity $\sqrt{q^*/r}$ can be interpreted as a ratio: $\sqrt{q^*/r} = \max\{(\beta + r)/2r, 2\}$ = strength of useful features in training/strength of those features in test. From equation (3.3), we learn that, in region I, the selected useful features perform about half as well in testing as we might expect from their performance in training.

4. Behaviour of ideal clipping threshold

This section sketches the arguments supporting theorems 3.2, 3.4, 3.6 and 3.7. In view of space restrictions, arguments supporting lemmas from this and later sections may be found in the electronic supplementary material.

In the RW model, it makes particular sense to use the clipping threshold function $\eta_{\text{clip}}^t$, since all non-zeros are known to have the same amplitude. The ideal clipping threshold is also very easy to analyse heuristically. However, it turns out that all the statements in theorems 3.2, 3.4, 3.6 and 3.7 are equally valid for all the three types of threshold function, so we prefer to explain the derivations using clipping.

(a) $\widetilde{\text{Sep}}$ in terms of true and false discoveries

In the RW model, we can express the components of the proxy separation very simply when using the clipping threshold:

$$A_{\text{clip}}(t, \epsilon, \tau) = \epsilon \tau E[\text{sgn}(\tau + W)1_{|\tau + W| > t}]$$

and

$$B_{\text{clip}}(t, \epsilon, \tau) = \epsilon E[1_{|\tau + W| > t}] + (1 - \epsilon) E[1_{|W| > t}],$$

where $W$ denotes an $N(0,1)$ random variable. Recall the definitions of useful selections TP and useless selections FP; we must also count inverted detections, for the case where the $\mu_i > 0$ but $\eta_{\text{clip}}^i(Z_i) < 0$. Put $E(\text{ID})(t; \epsilon, \tau, p) = \epsilon p\Phi(-t - \tau)$, with $\Phi$ again the standard normal distribution, and define the inverted detection rate by IDR = $p^{-1}E(\text{ID})$. Then

$$A_{\text{clip}}(t; \epsilon, \tau) = \tau (\text{TPR}(t) - 2\text{IDR}(t))$$

and

$$B_{\text{clip}}(t; \epsilon, \tau) = \text{TPR}(t) + \text{FPR}(t).$$

We arrive at an identity for $\widetilde{\text{Sep}}$ in the case of clipping:

$$\widetilde{\text{Sep}}(t; \epsilon, \tau) = 2\tau (\text{TPR}(t) - 2\text{IDR}(t))/\sqrt{\text{TPR}(t) + \text{FPR}(t)}.$$
To explain theorem 3.2, drop the term IDR and consider the alternative proxy
\[ \text{Sep}(t) = (2\tau)\text{TPR}(t)/\sqrt{\text{TPR}(t) + \text{FPR}(t)} = 2A(t)/B^{1/2}(t). \]

**Lemma 4.1.** Let \( \epsilon > 0 \) and \( \tau > 0 \). The threshold \( t_{\text{alt}} \) maximizing \( \text{Sep}(t) \) as a function of \( t \) satisfies the tangent–secant rule as an exact identity; at this threshold we have \( \text{Lfdr}(t_{\text{alt}}) = \frac{1}{2}(1 + \text{FDR}(t_{\text{alt}})) \).

The connection \( \text{Sep} \sim \text{alt} \) follows from part (iii) of the next lemma.

(b) Analysis in the asymptotic rare/weak model

We now invoke the ARW model: \( \epsilon = p^{-\beta}, \tau = \sqrt{2r \log p}, n \sim c(\log p)^r, p \to \infty \). Let \( t_p(q) = \sqrt{2q \log p} \). We have
\[ \Psi_0(t_p(q)) \sim \frac{\sqrt{2/\pi p^{-q}}}{t_p(q)}, \quad p \to \infty. \] (4.1)

We also need some notation for poly-log terms.

**Definition 4.2.** Any occurrence of the symbol \( \text{PL}(p) \) denotes a term that falls between \( C(\log p)^{-\zeta} \) and \( C(\log p)^{\zeta} \) as \( p \to \infty \) for some \( \zeta > 0 \) and \( C > 0 \). Different occurrences of this symbol may stand for different such terms.

In particular, we may well have \( T_1(p) = \text{PL}(p), T_2(p) = \text{PL}(p) \), as \( p \to \infty \), and yet \( T_1(p)/T_2(p) \not\to 1 \) as \( p \to \infty \). However, certainly \( T_1(p)/T_2(p) = \text{PL}(p), p \to \infty \).

The following lemma exposes the main phenomena driving theorems 3.2, 3.4, 3.6 and 3.7. It follows by simple algebra, and several uses of Mills’ ratio (4.1) in the convenient form \( \Psi_0(t_p(q)) = \text{PL}(p)p^{-\beta} \).

**Lemma 4.3.** In the asymptotic RW model, ARW, as \( p \to \infty \), we have the following results.

(i) Quasi power law for useful feature discoveries:
\[ E(\text{TP})(t_q(p), \epsilon, \tau) = \text{PL}(p)p^{\delta(q; \beta, \rho)}, \]
where the useful feature discovery exponent \( \delta \) obeys \( \delta(q; \beta, \rho) \equiv \begin{cases} 1 - \beta, & 0 < q \leq r, \\ 1 - \beta - (\sqrt{q} - \sqrt{r})^2, & r < q < 1. \end{cases} \)

(ii) Quasi power law for useless feature discoveries:
\[ E(\text{FP})(t_q(p), \epsilon, \tau) = \text{PL}(p)p^{1-q}. \]

(iii) Negligibility of inverted detections:
\[ E(\text{ID})(t_q(p), \epsilon, \tau) = o(E(\text{TP})(t_q(p), \epsilon, \tau)). \]

As an immediate corollary, under ARW, we have
\[ \text{Sep}(t_q(p), \epsilon, \tau) \left( \frac{1}{2}\sqrt{p/n} \right) = \text{PL}(p)p^{\delta(q; \beta, \rho)} / \sqrt{p^{\delta(q; \beta, \rho)} + p^{1-q}}, \quad p \to \infty. \]

On the right-hand side of this display, the poly-log term is relatively unimportant. The driving effect is the power-law behaviour of the fraction. The following lemma contains the core idea behind the appearance of \( p^* \) in theorems 3.2 and 3.4, and the distinction between region I and regions II and III.

*Phil. Trans. R. Soc. A (2009)*
Lemma 4.4. Let \((\beta, r) \in (\frac{1}{2}, 1)^2\). For fixed \(q, r\) and \(\beta\), let \(\gamma(q; r, \beta)\) denote the rate at which \(p^\beta(q; r, \beta) / (p^\rho(q; r, \beta) + p^{1-q})\) tends to \(\infty\) as \(p \to \infty\). Then \(\gamma > 0\) if and only if \(r > p^*(\beta)\). A choice of \(q\) maximizing this ratio is given by equation (3.3).

Lemma 4.4 shows that equation (3.3) gives us one choice of threshold maximizing the rate at which \((\sqrt{p/n} \text{Sep})\) tends to infinity. Is it the only choice? Define \(q_2 = q_2(r, \beta) = (\beta + r)^2 / (4r)\). Except in the case \(r > \beta\) (and so \(q_2 \in (\beta, r)\)), this is indeed the only choice. As the proof of lemma 4.4 shows, if \(r > \beta\), any \(q \in [\beta, r]\) optimizes the rate of separation. It turns out that, in that case, our formula \(q^*\) not only maximizes the rate of separation, but also correctly describes the leading-order asymptotic behaviour of \(T_{\text{ideal}}\). The key point is the tangent–secant formula, which picks out, from among all \(q \in [\beta, r]\), uniquely \(q_2\). This is shown by the next two lemmas, which thereby complete the proof of theorem 3.6.

Lemma 4.5. Set \(\gamma_0(q; r, \beta) = -\beta - r + 2\sqrt{rq}\), for \(q \in (0, 1)\). In the ARW\((r, \beta)\) model, consider the threshold \(t_q(p)\). Then, as \(p \to \infty\),

\[
\frac{\text{TPR}(t_q(p))}{\text{FPR}(t_q(p))} \sim \begin{cases} 
p^\gamma 2\sqrt{q - 2\sqrt{r}} & q > r, 
p^\beta \sqrt{t_q(p)} & q < r. \end{cases}
\]

Moreover, suppose \(q \neq r\). Then

\[
\frac{\text{TPR}'(t_q(p))}{\text{FPR}'(t_q(p))} = \frac{1}{2} p^\gamma, \quad p \to \infty.
\]

Lemma 4.6. The above defined \(q_2(r, \beta)\) is the unique solution of \(\gamma_0(q; r, \beta) = 0\). Suppose \(r > \beta\). As \(p \to \infty\), \(T_{\text{ideal}} / t_{q_2}(p) \to 1\).

(c) False discovery rate/local false discovery rate properties of ideal threshold

We now turn to Theorem 3.7 and make an important observation:

The asymptotic \(T_{\text{ideal}} = t_{q^*}(p)(1 + o(1))\) is simply too crude to determine the FDR and Lfdr properties of \(T_{\text{ideal}}\); it is necessary to consider the second-order effects implicit in the \(1 + o(1)\) term. For this, the tangent–secant formula is essential.

Indeed, equation (4.3) shows that the only possibilities for limiting local FDR of a threshold of the exact form \(t_q(p)\) are 0, 1/2, 1. The actual local FDR of \(T_{\text{ideal}}\) spans a continuum from \([0, 1]\), because small perturbations of \(t_q(p)(1 + o(1))\) implicit in the \(o(1)\) can cause a change in the local FDR. To understand this, for \(q \neq r\) and \(s \in (0, \infty)\), put

\[
\tilde{q}(q, r, s, p) = \left[ r^{1/2} \pm \sqrt{(q^{1/2} - r^{1/2})^2 + (\log s)/(\log p)} \right]^2,
\]

Phil. Trans. R. Soc. A (2009)
where the sign of ± is + if and only if \( q > r \). Clearly, \( \tilde{q} \) is well defined for all sufficiently large \( p \). As an example, \( \tilde{q}(q, r, 1, p) = q \). The peculiar definition ensures that \( \phi(t_q(p) - t_r(p)) = \phi(t_q(p) - t_r(p))s \).

By simple algebra one can show the following result.

**Lemma 4.7.** Let \( q, r \) and \( s \) be fixed. With \( \tilde{q} = \tilde{q}(q, r, s, p) \), \( t_q(p)/t_q(p) \to 1 \) and \( \phi(t_q(p))/\phi(t_q(p)) \to 1, p \to \infty \).

Let us put for short \( F_s = \text{FPR}(t_{\tilde{q}(q, r, s, p)}(p)) \) and \( T_s = \text{TPR}(t_{\tilde{q}(q, r, s, p)}(p)) \). Combining several identities from the last few sentences, we have, as \( p \to \infty \), \( T'_s = sT'_1 \), \( F'_s \sim F'_1 \), \( T'_s \sim sT_1 \) and \( F_s \sim F'_1 \). Hence,

\[
\begin{align*}
\text{FDR}(t_{\tilde{q}(q, r, s, p)}(p)) &= F_s/(T_s + F_s) \sim F_1/(sT_1 + F_1); \\
\text{Lfdr}(t_{\tilde{q}(q, r, s, p)}(p)) &= F'_s/(T'_s + F'_s) \sim F'_1/(sT'_1 + F'_1).
\end{align*}
\]

Choosing \( s \) appropriately, we can therefore obtain a perturbed \( q \) that perturbs the Lfdr and FDR. In fact, there is a unique choice of \( s \) needed to ensure the tangent–secant formula.

**Lemma 4.8.** For given values \( F_1, F'_1, T_1 \neq 0 \) and \( T'_1 \neq 0 \), put \( s^* = (F'_1/T'_1) - 2(F_1/T_1) \). This choice of \( s \) obeys the tangent–secant rule: \( F'_1/(s^* T'_1 + F'_1) = 1/2(1 + F_1/(s^* T_1 + F_1)) \).

To use this, recall equations (4.2) and (4.3). These formulae give expressions for \( T_1/F_1 \) and \( T'_1/F'_1 \). Plugging in \( q = q^* \), we get the following corollary.

**Corollary 4.9.** We have \( T_{\text{ideal}} \sim t_{\tilde{q}(q^*, r, s^*, p)}(p) \), where \( s^* \) is obtained by setting \( T_1 = (\sqrt{q^*} - \sqrt{r})^{-1} \), \( F_1 = 2/\sqrt{q} \), \( T'_1 = 1 \) and \( F'_1 = 2 \). Moreover, if \( \beta/3 < r < \beta \), \( \text{FDR}(T_{\text{ideal}}) \sim (\beta - r)/2r \), \( \text{Lfdr}(T_{\text{ideal}}) \sim (r + \beta)/4r \).

5. Connection of higher criticism objective with \( \widetilde{\text{Sep}} \)

Let \( F = F_{\epsilon, \tau} \) be the two-point mixture of definition 2.3 and \( G = G_{\epsilon, \tau} \) the corresponding folded distribution. For \( t = t_{\tilde{q}}(p) \) in the asymptotic RW model, we have

\[
\begin{align*}
\text{HC}(t; F_{\epsilon, \tau}) &= G(t) - \Psi(t) \\
&= \frac{\epsilon(\Psi_\tau - \Psi_0)(t)}{\sqrt{G(t)G(t)}} \\
&= \frac{\epsilon(\Psi_\tau - \Psi_0)(t)}{\sqrt{[((1 - \epsilon)\Psi_0 + \epsilon\Psi_\tau)((1 - \epsilon)\Psi_0 + \epsilon\Psi_\tau)]^{1/2}}} \\
&\sim \frac{\epsilon(\Psi_\tau - \Psi_0)(t)}{\sqrt{[((1 - \epsilon)\Psi_0 + \epsilon\Psi_\tau)]^{1/2}}} \\
&\sim \frac{\text{TPR}(t) - \epsilon/(1 - \epsilon)\text{FPR}(t)}{\text{TPR}(t)} \\
&\sim \frac{\text{TPR}(t) + \text{FPR}(t)}{\text{TPR}(t)} \\
&\sim \frac{(1 + o(1))\widetilde{\text{Sep}}(t; \epsilon, \tau), \quad p \to \infty.}
\end{align*}
\]
Table 2. Comparison of the exponents in Sep.

<table>
<thead>
<tr>
<th>method</th>
<th>Sep exponent</th>
<th>boundary, $3/4 &lt; \beta$</th>
<th>boundary, $1/2 &lt; \beta &lt; 3/4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ideal, HCT</td>
<td>$\gamma$</td>
<td>$1 - \sqrt{1 - \beta}$</td>
<td>$\beta - 1/2$</td>
</tr>
<tr>
<td>FDRT</td>
<td>$\frac{1}{2} - \frac{(\beta + r)^2}{8r}$</td>
<td>$1 - \sqrt{1 - \beta}$</td>
<td>$1 - \sqrt{1 - \beta}$</td>
</tr>
<tr>
<td>Bonferroni</td>
<td>$-\beta - r + 2\sqrt{r}$</td>
<td>$1 - \sqrt{1 - \beta}$</td>
<td>$1 - \sqrt{1 - \beta}$</td>
</tr>
</tbody>
</table>

Step (5.1)$_3$ follows from $G_{\epsilon, \tau}(t_q(p)) \to 1$ as $p \to \infty$. Step (5.1)$_5$ follows from $\epsilon \text{FPR}(t_q(p)) = o(\text{TPR}(t_q(p)))$ as $p \to \infty$. Step (5.1)$_6$ follows from $\text{TPR}(t_q(p)) = o(\text{FPR}(t_q(p)))$ as $p \to \infty$.

This derivation can be made rigorously correct when $q = q^*(r, \beta)$, where $q^*$ is as announced in theorem 3.4. With extra work, not shown here, one obtains the following theorem.

**Theorem 5.1.** The statements made for the ideal threshold in theorems 3.2, 3.4 and 3.6 are equally valid for the ideal HC threshold.

6. Suboptimality of phase diagram for other methods

Thresholds can instead be chosen by false discovery rate control (FDRT) and by control of the family-wise error rate (FWERT), or Bonferroni. Theorems 3.4 and 3.6 implicitly show the suboptimality of these approaches. The electronic supplementary material sketches an argument implying the following theorem.

**Theorem 6.1.** Under the asymptotic RW model $\text{ARW}(r, \beta)$, the regions of the $(r, \beta)$ phase space for successful classification by FDRT and FWERT are $r > (1 - \sqrt{1 - \beta})^2$, $0 < \beta < 1$.

**Remark.** The success regions for FDRT and FWERT are smaller than for HCT. However, even over the region of joint success, HCT still yields asymptotically much larger Fisher separations than the other principles. Figure 3 displays the exponents of the Fisher separations for $\beta = 1/2$ and $\beta = 5/8$ (see also table 2).

7. Discussion and conclusions

These results form part of a series including Donoho & Jin (2008, in preparation) and Jin (2009). Donoho & Jin (2008) reported numerical results both with simulated data from the RW model and with real datasets often used as standard benchmarks for classifier performance. The current paper develops an analysis of ideal HCT that is both transparent and insightful. While this paper focuses on the case where $n \sim c(\log p)^\gamma$, the more general cases where $n$ is fixed as $p$ grows and $n = n_p = p^\theta$ for some $\theta \in (0, 1)$ require much new discussion. This wider setting is considered in Jin (2009) (see also Ingster et al. in press), who proved the impossibility of successfully training classifiers in what we call the region of...
impossibility. In Donoho & Jin (in preparation), we will develop a complete set of results for the more general linkages of $n$ with $p$ and for HCT as opposed to ideal HCT.

Beyond ideal performance. Conceptually, the ideal threshold envisages a situation with an oracle, who, knowing $\epsilon$ and $\tau$, and $n$ and $p$, chooses the very best threshold possible under those given parameters. In this paper we have analysed the behaviour of this threshold within a certain asymptotic framework. Clearly, no empirical procedure can duplicate the performance of the ideal threshold.

However, HC thresholding comes close. The HC threshold does not involve optimal exploitation of knowledge of $\epsilon$ and $\tau$; it needs only the feature $Z$-scores. In this paper, we analysed the behaviour of the ideal HCT: $t_{HC} = T_{HC}(F_{t,\tau}).$

This is an ideal procedure because in a real-life situation we do not know the distribution $F_{t,\tau}$ of $Z$-scores; we have instead the empirical cumulative distribution function (CDF) $F_{n,p}$ defined by $F_{n,p}(z) = (1/p) \sum_{j=1}^{p} 1\{Z(j) \leq z\}$. The (non-ideal) HCT that we defined in §1 is then simply $t_{HC}^{n,p} = T_{HC}(F_{n,p})$. Because $F_{t,\tau}(z) = E(F_{n,p})(z)$, there is nothing controversial about the assertion that $t_{HC} \approx t_{HC}^{n,p}$ for large $n$ and $p$. Indeed, there are generations of experience for other functionals $T$ showing that we typically have $T(F_{n,p}) \approx T(E(F_{n,p}))$ for large $n$, $p$ for those other functionals. Proving this for $T = T_{HC}$ is more challenging than one might anticipate; the problem is that the functional $T_{HC}$ is not continuous at $F = \Phi$, and yet $F_{t,\tau} \to \Phi$ as $p \to \infty$. After considerable effort, we justify the approximation of HCT by ideal HCT in Donoho & Jin (in preparation). The analysis presented here only partially proves that HCT gives near-optimal threshold feature selection; it only explains the connection at the level of ideal quantities. But this is the main insight.

Phase diagram for finite sample sizes. The phase diagram correctly reflects finite $(n,p)$ behaviour. In figure 4, we consider $p = 3 \times 10^3 \times N$, $N = (1, 10, 100)$. For such $p$, we take $n = (\log p)/2$ and display the boundary of the set of $(\beta, r)$, where ideal HCT yields a classification error between 10 and 40 per cent. As $p$ grows, the upper and lower bounds migrate towards the limit curve $r = \rho^{*}(\beta)$.

Other work on HC. HC was originally proposed for use in a detection problem that has nothing to do with threshold feature selection: testing an intersection null hypothesis $\mu(j) = 0$, for all $j$ (Donoho & Jin 2004). The literature has developed since then. Hall & Jin (2008, 2009) extend the optimality of HC in detection to correlated settings. Wellner and his collaborators investigated HC in the context of goodness of fit; see for example Jager & Wellner (2007). Hall and his collaborators have investigated HC for robustness; see for example Delaigle & Hall (2009). HC has been applied to data analysis in cosmology and astronomy (Cayon et al. 2005; Jin et al. 2005). HC has also been used as a principle to synthesize new procedures: Cai et al. (2007) use HC to motivate estimators for the proportion $\epsilon$ of non-null effects.

HC in classification. HC was previously applied to high-dimensional classification in Hall et al. (2008), but with a key conceptual difference from the present paper. Our approach uses HC in classifier training—it selects features in designing a linear classifier—but the actual classification decisions are made by the linear classifier when presented with specific test feature vectors.
Figure 3. Comparison of Sep exponents for ideal HCT (red), FDRT (green) and Bonferroni (light blue): (a) $\beta = 1/2$; (b) $\beta = 5/8$. The x-axes display r-values from 0 to $\beta$; y-axes display corresponding exponents. The solid blue horizontal bar in (b) indicates that r falls in the failure region where all exponents are 0.

Figure 4. Classification boundaries for $p = 3 \times 10^3 \times (1,10,100)$. For each $p$, both an upper (U) boundary and a lower (L) boundary are calculated, which correspond to the proxy classification errors of 10 and 40 per cent using ideal HCT. The gap between the two boundaries gets smaller as $p$ increases. The solid black curve displays the asymptotic boundary $r = \rho^\ast(\beta)$. Lines with (without) dots represent lower (upper) boundaries: $p = 3000$, red; 30 000, green; 300 000, blue.

The approach in Hall et al. (2008) uses HC to directly make decisions from feature vector data. Let us call these two different strategies HC-based feature selection (used here) and HC-based decision (used in Hall et al. 2008).
In the ARW model, for a given classifier performance level, the HC-based decision strategy requires much stronger feature contrasts to reach that level than the HC-based feature selection strategy. In this paper, we have shown that HC feature selection requires useful features to have contrasts exceeding \( \sqrt{2\rho^*(\beta)\log p} / \sqrt{n(1 + o(1))} \), \((n, p) \to \infty\). HC decision requires useful features to have contrasts exceeding \( \sqrt{2\rho^*(\beta)\log p(1 + o(1))} \), \(p \to \infty\). Therefore, if the number of training samples \( n > 1 \), HC feature selection has an advantage. Imagine that we have \( n = 36 \) training samples; we will find that HC feature selection can asymptotically detect features at roughly \( 1/6 \) the strength required when using HC directly for decision.

Other work on the RW model. During the review of this paper, a new manuscript by Ingster et al. (2009) appeared in arXiv analysing a feature selection setting equivalent to the ARW model, using non-HC-based approaches.

The authors would like to thank the Isaac Newton Mathematical Institute at Cambridge for hospitality during the program Statistical Theory and Methods for Complex, High-Dimensional Data, and for a Rothschild Visiting Professorship held by D.D. We would also like to thank an anonymous referee for many helpful comments. D.D. was partially supported by NSF DMS-0505303, and J.J. was partially supported by NSF CAREER award DMS-0908613.

References


Donoho, D. & Jin, J. In preparation. If useful features are rare and weak, HCT yields successful classification where success is possible.

Phil. Trans. R. Soc. A (2009)


