Higher Criticism Thresholding: Optimal Feature Selection when Useful Features are Rare and Weak

David Donoho [*] **and Jiashun Jin** [†] [‡]

[*]Stanford University, Stanford, CA 94305,[†]Purdue University, West Lafayette, IN 47907, and [‡]Carnegie Mellon University, Pittsburgh, PA 15213

Contributed by David L. Donoho, July 30, 2008

Linear classification analysis is a fundamental tool for science and technology. In important application fields today – genomics and proteomics are examples – one automatically obtains very high-dimensional feature vectors; but in a specific project, few of those routinely measured features will be useful. The subset of useful features is unknown at the outset of a new project and must be learned from a training set; the task is complicated by the typical small size of the training set, the noisiness of the data, and the high dimensionality of the feature set.

We study feature selection by thresholding of feature Z-scores and introduce a new principle of threshold selection, based on the notion of Higher Criticism (HC). For $i = 1, 2, \dots, p$, let π_i denote the two-sided P-value associated to the i-th feature Z-score and $\pi_{(i)}$ denote the i-th order statistic of the collection of P-values. The HC threshold is the absolute Z-score corresponding to the P-value maximizing the HC objective $(i/p - \pi_{(i)})/\sqrt{i/p(1 - i/p)}$.

We consider a Rare/Weak (RW) feature model, where the fraction of useful features is small and the useful features are each too weak to be of much use on their own. HC thresholding (HCT) has interesting behavior in this setting, with an intimate link between maximizing the HC objective and minimizing the error rate of the designed classifier, and very different behavior from popular threshold selection procedures such as false-discovery rate thresholding (FDRT).

In the most challenging RW settings, HCT uses an unconventionally low threshold; this keeps the missed-feature detection rate under better control than FDRT and yields a classifier with improved misclassification performance but a false feature selection rate very much higher than the current fashion: in some settings a very high fraction of the features selected by HCT will actually be useless. In the case of many features and few observations, cross-validated threshold selection (CVT) is known to be highly variable; in the RW model, replacing CVT in the popular Shrunken Centroid classifier with the computationally less expensive and simpler HCT reduces the variance of the selected threshold and the error rate of the constructed classifier. Results on standard real datasets and in asymptotic theory confirm the advantages of HCT.

Linear Classification | Feature Selection | Threshold Selection | Higher Criticism | False Discovery Rate | Shrunken Centroids.

Introduction

The modern era of high-throughput data collection creates data in abundance; however, this data glut poses new challenges. Consider a simple model of linear classifier training. We have a set of labelled training samples (Y_i, X_i), $i = 1, \dots, n$, where each label Y_i is ± 1 and each feature vector $X_i \in R^p$. For simplicity, we assume the training set contains equal numbers of 1's and -1's and that the feature vectors $X_i \in R^p$ obey $X_i \sim N(Y_i\mu, \Sigma)$, $i = 1, \dots, n$, for an unknown mean contrast vector $\mu \in R^p$; here Σ denotes the feature covariance matrix and n is the training set size. In this simple setting, one ordinarily uses linear classifiers, taking the general form $L(X) = \sum_{j=1}^p w(j)X(j)$, for a sequence of 'feature weights' $w = (w(j) : j = 1, \dots, p)$.

Classical theory going back to RA Fisher [4] shows that the optimal classifier has feature weights $w \propto \Sigma^{-1}\mu$; at first glance linear classifier design seems straightforward and settled. However, in many of today's most active application areas, it is a major challenge to construct linear classifiers which work well.

In many ambitious modern applications – genomics and proteomics come to mind – measurements are automatically made on thousands of standard features, but in a given project, the number of observations, n, might be in the dozens or hundreds. In such settings, $p \gg n$, which makes it difficult or impossible to estimate the feature covariance straightforwardly. In such settings one often ignores feature covariances. Working in standardized feature space where individual features have mean zero and variance one, a by-now standard choice uses weights $w(j) \propto Cov(Y, X(j)) \equiv \mu(j)$ [6, 17]. Even when this reduction makes sense, further challenges remain.

When Useful Features are Rare and Weak. In some important applications, standard measurements generate many features automatically, few of which are likely to be useful in any specific project, but researchers don't know in advance which ones will be useful in a given project. Moreover, reported misclassification rates are relatively high. Hence the dimension p of the feature vector is very large, and although there may be numerous useful

---

Reserved for Publication Footnotes

features, they are relatively rare and individually quite weak.

Consider the following *rare/weak feature model (RW Feature Model).* We suppose the contrast vector μ to be nonzero in only k out of p elements, where $\epsilon = k/p$ is small, i.e. close to zero. As an example, we might have $p = 10,000$, $k = 100$, and so $\epsilon = k/p = .01$. In addition, we suppose that the nonzero elements of μ have *common* amplitude μ_0. Since the elements $X(j)$ of the feature vector where $\mu(j) = 0$ 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. Denote $\tau = \sqrt{n}\mu_0$, we speak of the parameters ϵ and τ as the sparsity and strength parameters and denote by $RW(\epsilon, \tau)$ this setting. (Related 'sparsity' models are common in estimation settings [13, 14]. The RW model includes an additional feature strength parameter τ not present in those estimation models. More closely related to the RW model is work in multiple testing by Ingster and the authors [22, 11, 23] although the classification setting gives it a different meaning).

Naive application of the formula $w \propto Cov(Y, X)$ in the RW setting often leads to very poor results; the vector of empirical covariances $(\widehat{Cov}_{n,p}(Y, X(j)) : j = 1, \ldots, p)$ is very high-dimensional and contains mostly 'noise' coordinates; the resulting naive classification weights $\hat{w}_{naive}(j) \propto \widehat{Cov}_{n,p}(Y, X(j))$ often produce correspondingly noisy decisions. The data glut seriously damages the applicability of such 'textboook' approaches.

Feature Selection by Thresholding. Feature selection - - i.e. working only with an empirically-selected subset of features - is a standard response to data glut. Here and below, we suppose that feature correlations can be ignored and that features are standardized to variance one. We consider subset selectors based 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 $H_{0,j}$: $Cov(Y, X(j)) = 0$. Under our assumptions $Z \sim N(\theta, I_p)$ where $\theta = \sqrt{n}\mu$ and μ is the feature contrast vector. Features with nonzero $\mu(j)$ typically have significantly nonzero $Z(j)$ and conversely, other features will have $Z(j)$ 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: $\eta_t^\star(z)$, $\star \in \{clip, hard, soft\}$. These are: *Clipping* – $\eta_t^{clip}(z) = \mathrm{sgn}(z)$, which ignores the size of the Z-score, provided it is large; *Hard Thresholding* – $\eta_t^{hard}(z) = z \cdot 1_{\{|z|>t\}}$, which uses the size of the Z-score, provided it is large; and *Soft Thresholding* – $\eta_t^{soft}(z) = \mathrm{sgn}(z)(|z| - t)_+$, which uses a shrunken Z-score, provided it is large.

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

In words, the classifier sums across features with large training-set Z-scores, using a simple function of the Z-score to weight the corresponding feature appropriately. Well-known methods for linear classification follow this approach: the Shrunken Centroids method [25] reduces in our two-class setting to a variant of soft thresholding; the schemes discussed in [18] and [21] are variants of hard thresholding.

Thresholding has been popular in estimation for more than a decade [13]; it is known to be succesful in 'sparse' settings where the estimand has many coordinates, of which only a relatively few coordinates are significantly nonzero. While classification is not the same as estimation, an appropriate theory for thresholding can be constructed [12] showing that threshold feature classifiers with ideally-chosen thresholds work well and even optimally control the misclassification rate.

One crucial question remains: how to choose the threshold based on the data? Related proposals for threshold choice include cross-validation [25]; control of the false discovery rate [2, 1]; and control of the local false discovery rate [16].

Higher Criticism

We propose a method of threshold choice based on recent work in the field of multiple comparisons.

HC Testing. Suppose we have a collection of N P-values π_i which under the global null hypothesis are uniformly distributed: $\pi_i \sim_{iid} U[0, 1]$. We perform the increasing rearrangement into order statistics: $\pi_{(1)} \le \pi_{(2)} \le \cdots \le \pi_{(N)}$; and we note that, under the null hypothesis, these order statistics have the usual properties of uniform order statistics, including the asymptotic normality $\pi_{(i)} \sim_{approx} \mathrm{Normal}(i/N, i/N(1 - i/N))$. The ordered P-values may be compared with such properties, leading to the following notion:

Definition 2. *(**HC Testing**) [11] The Higher Criticism objective is*

$$HC(i; \pi_{(i)}) = \sqrt{N} \frac{i/N - \pi_{(i)}}{\sqrt{i/N(1 - i/N)}}. \quad \textbf{[1]}$$

Fix $\alpha_0 \in (0, 1)$ (eg $\alpha_0 = 1/10$). The HC test statistic is $HC^* = \max_{1 \le i \le \alpha_0 N} HC(i; \pi_{(i)})$.

In practice, HC is typically insensitive to the selection of α, especially in Rare/Weak situations. The HC-objective function is the "Z-score of the P-value", i.e. a standardized quantity with asymptotic distribution $N(0, 1)$ under the null hypothesis. In words, we look for the largest standardized discrepancy between the expected behavior of the π_i under uniformity and the observed behavior. When this is large, the whole collection of P-values is not consistent with the global null hypothesis. The phrase "Higher Criticism" reflects the shift in emphasis from single test results to the whole collection of tests [11]. The HC test statistic was developed to detect the presence of a small fraction of non null hypotheses among many truly null hypotheses [11]. Note: there are several variants of HC statistic; we discuss only one variant in this brief note; the main results of [11] still apply to this variant; for full discussion see [11, 12].

HC Thresholding. Return to the classification setting in previous sections. We have a vector of feature Z-scores $(Z(j), j = 1, \ldots, p)$. We apply HC notions by translating the Z-scores into two-sided P-values, and maximizing the HC objective over index i in the appropriate range.

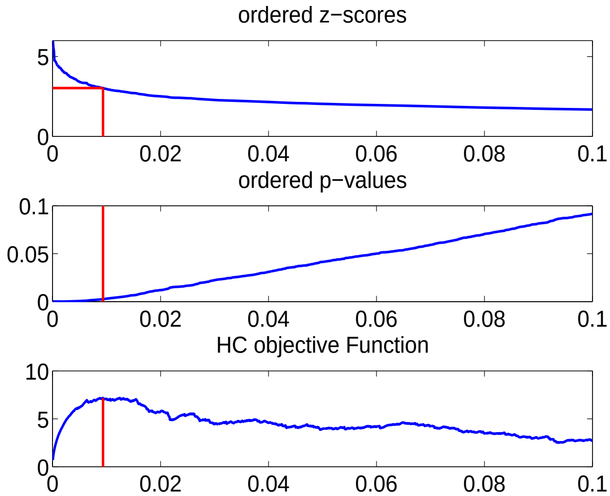

Fig. 1. Illustration of HC Thresholding. Panel (a) the ordered $|Z|$-scores. Panel (b) the corresponding ordered P-values in a PP plot. Panel (c) the HC objective function in Eq. 1; this is largest at $\hat{i} \approx 0.01N$ (x-axes are i/N). Vertical lines indicate $\pi_{(\hat{i})}$ in Panel (b), and $|Z|_{(\hat{i})}$ in Panel (a).

Mixing standard HC notations with standard multivariate data notation requires a bit of care. Please recall that p always refers to the number of measured classifier features while terms such as "P-value" and "$\pi_{(i)}$" refer to unrelated concepts in the HC setting. In an attempt to avoid notational confusion let $N \equiv p$ and sometimes use N in place of p. Define the *feature P-values* $\pi_i = Prob\{|N(0,1)| > |Z(i)|\}$, $i = 1, \dots, N$; 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 3. *(**HC Thresholding.**) Apply the HC test to the feature P-values. Let the maximum HC objective be achieved at index $\hat{i}$. The **Higher Criticism threshold** (HCT) is the value $\hat{t}^{HC} = |Z|_{(\hat{i})}$. The **HC threshold feature selector** selects features with Z-scores exceeding $\hat{t}^{HC}$ in magnitude.*

Figure 1, Panels a-c, illustrates the procedure. Panel a shows a sample of Z-scores, Panel b shows a PP-plot of the corresponding ordered P-values versus i/N and Panel c shows a standardized PP-plot. The standardized PP-Plot has its largest deviation from zero at $\hat{i}$; and this generates the threshold value.

Performance of HCT in RW feature model

In the $RW(\epsilon, \tau)$ model, the feature Z-scores vector $Z \sim N(\theta, I_p)$, where θ is a sparse vector with fraction ϵ of entries all equal to τ and all other entries equal to 0.

Figure 2, panels a-d, exhibits results from a collection of problems all with $p = 1000$ features, of which only 50 are truly useful – i.e. have θ nonzero in that coordinate – so that in each case the fraction of useful features is $\epsilon = 50/1000 = 0.05$. In this collection, the amplitude τ of nonzeros varies from 1 to 3. Here the useful features are indeed weak: they have expected Z-scores typically lower than some Z-scores of useless coordinates!

We compare HC thresholding with 3 other thresholding rules: (a) FDRT(.5) – thresholding with false feature discovery rate control parameter $q = 0.5$, (b) FDRT(.1) – thresholding with false feature discovery rate (FDR) control parameter $q = 0.1$, and (c) Bonferroni – setting the threshold so that the expected number of false

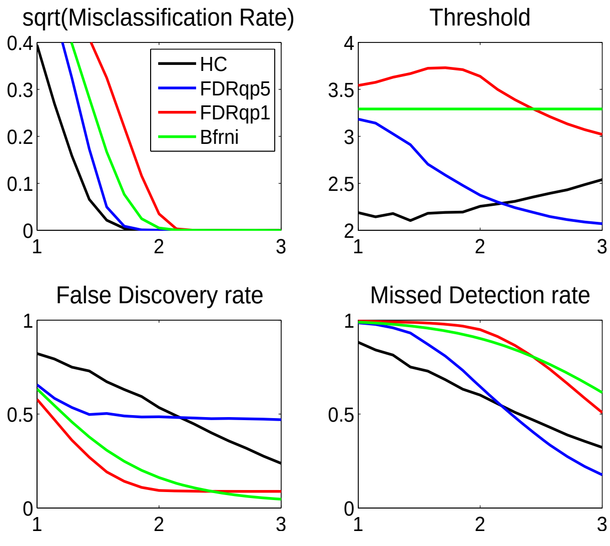

Fig. 2. Monte-Carlo Performance of Thresholding Rules in the RW model. In all the panels, $p = 1000$, $\epsilon = 0.05$, and x-axes display τ. Upper Left: MCR$^{1/2}$. Upper Right: Average Threshold. Lower left: Average FDR. Lower Right: Average MDR. Threshold procedures used: HC (black), Bonferroni (Green), FDR ($q = .5$) (Blue), FDRT ($q = .1$) (Red). Averages from 1000 Monte-Carlo realizations.

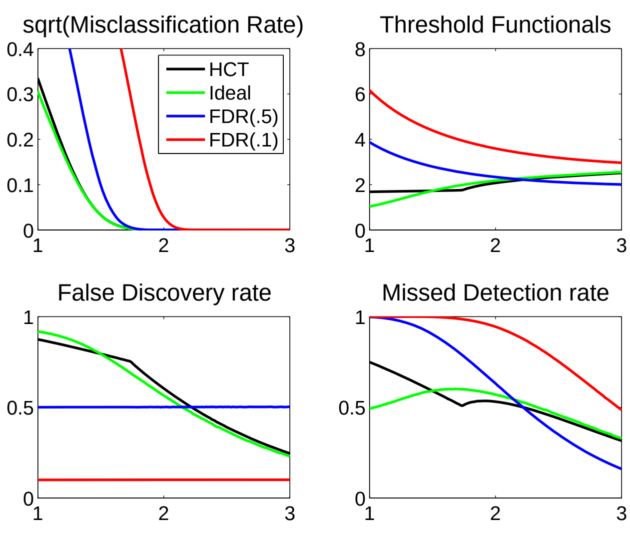

Fig. 3. Comparison of HCT Functional with Ideal Functional. In all the panels, $\epsilon = 1/100$, and x-axes display τ. Upper Left: MCR$^{1/2}$. Upper Right: Threshold. Lower Left: FDR. Lower Right: MDR. Threshold procedures used: HC (black), Ideal (green). Curves for FDR thresholding with $q = .5$ (blue) and $q = .1$ (red) are also shown. In each measure, green and black curves are close for $\tau > 2$. The discrepancy at small τ is caused by the limitation $T_{HC} > t_0$.

features is 1. These three rules illustrate what we believe to be today's orthodox opinion, which strives to ensure that most features in the classification rule are truly useful, and to strictly control the number of useless features present in the trained classifier. Local false discovery rate control shares the same philosophy. We generated 1000 monte-carlo realizations at each choice of parameters. We present results in terms of the dimensionless parameter τ, which is independent of n; if desired the reader may choose to translate these results into the form $\mu_0 = \tau/\sqrt{n}$ for a conventional choice of n, such as $n = 40$. Figure 2 presents the empirical average performance. As compared to traditional approaches, HCT has, in the case of weak signals, a lower threshold, a higher false feature discovery rate and lower missed feature detection rate (MDR); the misclassification rate (MCR) is also improved. In these displays, as the signal strength τ increases, HCT increases but FDRT decreases. For analysis of this phenomenon, see [12].

The HCT Functional and Ideal Thresholding

We now develop connections between HCT and other important notions.

HCT Functional. The *HCT functional* is, informally, the "threshold that HCT is trying to estimate". More precisely, note that, in the $RW(\epsilon,\tau)$ model, the empirical distribution function $F_{n,p}$ of feature Z-scores $F_{n,p}(t) = Ave_j I_{\{Z(j)\le t\}}$, approximates, for large p and n arbitrary, the theoretical CDF $F_{\epsilon,\tau}(t) = (1-\epsilon)\Phi(t) + \epsilon\Phi(t-\tau)$, $t \in \mathbf{R}$, where $\Phi(t) = P\{N(0,1) \le t\}$ is the standard normal distribution. The HCT functional is the *result of the HCT recipe upon systematically replacing $F_{n,p}(t)$ by $F_{\epsilon,\tau}(t)$*.

We define the underlying *True Positive Rate* $TPR(t)$, the *False Positive Rate* $FPR(t)$, and the positive rate $PR(t)$ in the natural way as the expected proportions of, respectively, the useful, the useless, and of all features, having Z-scores above threshold t. The HC objective functional can be rewritten (up to rescaling) as

$$\widetilde{HC} = \frac{PR(t) - FPR(t)}{\sqrt{PR(t)(1-PR(t))}} = \frac{\epsilon(TPR(t) - FPR(t))}{\sqrt{PR(t)(1-PR(t))}}. \qquad [2]$$

In the $RW(\epsilon,\tau)$ model, we have $TPR(t;\epsilon,\tau) = \Phi(t-\tau) + \Phi(-t-\tau)$, $FPR(t;\epsilon,\tau) = 2\Phi(-t)$, and $PR(t;\epsilon,\tau) = (1-\epsilon)FPR(t) + \epsilon\cdot TPR(t)$. Let $t_0 = t_0(\epsilon,\tau)$ denote the threshold corresponding to the maximization limit α_0 in Definition 2: $PR(t_0;\epsilon,\tau) = \alpha_0$. The HCT functional solves a simple maximization in t

$$T_{HC}(F_{\epsilon,\tau}) = \text{argmax}_{t\ge t_0}\widetilde{HC}(t;\epsilon,\tau). \qquad [3]$$

Rigorous justification of this formula is supplied in [12], showing that in the $RW(\epsilon,\tau)$ model, $\hat{t}^{HC}_{n,p}$ converges in probability to $T_{HC}(F_{\epsilon,\tau})$ as p goes to infinity with n either fixed or increasing; so indeed, this is what *HCT* is 'trying to estimate'.

Ideal Threshold. We now study the threshold which, (if we only knew it!) would provide optimal classifier performance. Recall that in our setting, the feature covariance is the identity $\Sigma = I_p$; the quantity $Sep(w;\mu) = w'\mu/\|w\|_2$ is a fundamental measure of linear classifier performance. The misclassification rate of the trained weights $\hat{w}$ on independent test data with true contrast vector μ obeys

$$P\{\text{Error}|\text{Training Data}, \mu\} = \Phi(-Sep(\hat{w};\mu)), \qquad [4]$$

where again Φ is the standard normal $N(0,1)$ CDF. Hence maximizing Sep is a proxy for minimizing misclassification rate. (For more details, see [12]).

For a fixed threshold t, let $Sep(w^t_{clip};\mu)$ denote the realized value of Sep on a specific realization. For large p and n arbitrary, this is approximately proportional to

$$\widetilde{Sep}(t) = \frac{(TPR(t) - 2\cdot TSER(t))}{\sqrt{PR(t)}}, \qquad [5]$$

where TPR and PR are the true and positive rates defined earlier, and $TSER(t)$ denotes the expected *True Sign Error Rate* $TSER(t) \equiv P\{Z(j) < 0|\mu(j) > 0\}$. We are in the RW model, so $\widetilde{Sep}(t;\epsilon,\tau)$ can be written in terms of $TPR(t;\epsilon,\tau)$, $PR(\epsilon,\tau)$, and $TSER(t;\tau) = \Phi(-t-\tau)$. We define the ideal threshold functional

$$T_{Ideal}(F_{\epsilon,\tau}) = \text{argmax}_t\widetilde{Sep}(t;\epsilon,\tau). \qquad [6]$$

Among all fixed thresholds, it achieves the largest separation for a given underlying instance of the $RW(\epsilon,\tau)$ model.

Comparison. How does the HCT functional compare to the ideal threshold functional, both in value and performance? They seem surprisingly close. Figure 3, panels a-d, presents the values, FDR, MDR, and MCR for these functionals in cases with $\epsilon > 0$ fixed and τ varying. The HCT functional quite closely approximates the ideal threshold, both in threshold values and in performance measures. In particular, we note that the behavior of the HCT rule that was so distinctive in the Rare/Weak

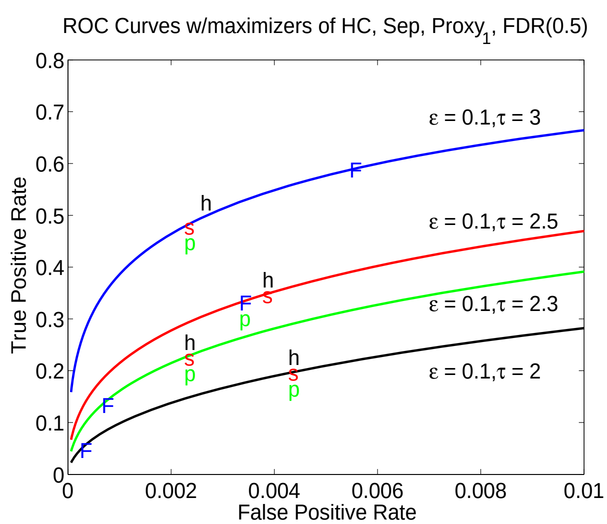

Fig. 4. Receiver Operating Characteristics curves for threshold detectors, together with operating points of (h) max-HCT, (s) max-SEP, (p) max-$Proxy_1$. Also included are the operating points of (F) FDR thresholding with $q = .5$. Note that h,s, and p are quite close to each other, while F can be very different.

features model – high false feature discovery rate and controlled missed detection rate – are actually behaviors seen in the ideal threshold classifier as well. The systematic discrepancy between the HCT and the ideal threshold at small τ is due to the constraint $t > t_0$ in Eq. 3.

ROC Analysis. The similarity between the HCT functional and the ideal threshold functional derives from the expressions for $PR(t;\epsilon,\tau)$ and $TPR(t;\epsilon,\tau)$ in the RW model. In the setting where very few features are selected, $(1 - PR(t)) \approx 1$, and $2TSER(t) \ll TPR(t)$, so we see by comparison of Eq. 2 and Eq. 5 that $\widetilde{HC}(t;\epsilon,\tau) \approx \widetilde{Sep}(t;\epsilon,\tau)$, as reflected in Figure 3 (For more discussion and analysis, see [12]).

Consider two related expressions: $Proxy_1 = \epsilon \cdot TPR(t)/\sqrt{PR(t)}$, $Proxy_2 = \epsilon \cdot TPR(t)/\sqrt{FPR(t)}$. Maximizing either of these proxies over t is equivalent to seeking a certain point on the so called *Receiver-Operating Characteristics* (ROC) curve $(FPR(t), TPR(t) : 0 < t < \infty)$. Figure 4 shows a range of ROC curves; the maximizers of $\widetilde{HC}(t)$, of $\widetilde{Sep}(t)$ and of $Proxy_1(t)$ are very close in the ROC space. Since the misclassification rate $MCR(t) = (1-\epsilon)(1 - FPR)(t) + \epsilon \cdot (1 - TPR)(t)$ is a Lipschitz function of the ROC coordinates, all three maximizers must offer similar performance.

The maximizer of $Proxy_2$ has a very elegant characterization, as the point in t where the *secant* to the ROC curve is double the *tangent* to the ROC curve, $\frac{TPR'}{FPR'} = \frac{TPR}{2FPR}$ at $t = t_{Proxy_2}$. The maximizer of $Proxy_1$ obeys a slightly more complex relationship $\frac{TPR'}{FPR'} = (2\frac{FPR}{TPR}(1 - \epsilon/2)(1-\epsilon)+\epsilon)^{-1}$ at $t = t_{Proxy_1}$. For small enough ϵ, this nearly follows the same rule: secant $\approx 2\times$ tangent rule.

For comparative purposes, FDR thresholding finds a point on the ROC curve with prescribed *secant*: $\frac{TPR}{FPR} = \frac{1-\epsilon}{\epsilon}(q^{-1} - 1)$ at $t = t_{FDR,q}$. Further, a *local* false discovery rate threshold yields a point on the ROC curve with prescribed *tangent* $\frac{TPR'}{FPR'} = \frac{1-\epsilon}{\epsilon}(q^{-1} - 1)$ at $t = t_{localFDR,q}$. Defining the *True Discovery Rate* $TDR \equiv 1 - FDR$, we see that HCT obeys $\frac{FDR}{TDR} \approx \frac{1}{2}\frac{localFDR}{localTDR}$, at $t = t_{Proxy_2}$. HCT and its proxies are thus visibly quite different from prescribing FDR or local FDR, which again underscores the distinction between avoiding false feature selection and maximizing classifier performance.

Complements

Performance on Standard Datasets. In recent literature on classification methodology, a collection of 6 datasets has been frequently used for illustrating empirical classifier performance [9]. We have reservations about the use of such data to illustrate HCT, because no one can say whether any specific such dataset is an example of Rare/Weak feature model. However, such comparisons are sure to be requested, so we report them here.

Of the standard datasets reported in [9], three involve two-class problems of the kind considered here; these are the `ALL` [18], `Colon` [3], and `Prostate` [24] datasets. In [9], 3-fold random training-test splits of these datasets were considered, and 7 well-known classification procedures were implemented: Bagboost [9], LogitBoost [10], SVM [8], Random Forests [7], PAM [25], and the classical methods *DLDA*, *KNN*. We applied HCT in a completely out-of-the-box way using definitions standard in the literature (for full details, see [20]). HCT-hard, which uses feature weights based on hard thresholding of feature Z-scores, gave quite acceptable performance. For comparison, introduce the relative regret measure $Regret(A) = [err(A) - min_{A'} err(A')]/[max_{A'} err(A') - min_{A'} err(A')]$. This compares the achieved error rate with the best and worst performance seen across algorithms. We report errors rates and regrets side by side in Table 1, where rows 2–7 are from Dettling [9], row 8 is provided by Tibshirani, and row 9 is the result of HCT-hard. Additionally, column 5 is the maximum regret across 3 different data sets, column 6 is the rank based on the maximum regret. In the random-split test, HCT-hard was the minimax regret procedure, always being within 29% of the best known performance, while every other procedure was worse in relative performance in at least some cases.

It is worth remarking that HCT-based feature selection classifiers are radically simpler than all the other methods being considered in this competition, requiring no tuning or cross-validation to achieve the presented results.

Comparison to Shrunken Centroids. The well known 'Shrunken Centroids' (SC) algorithm [25] bears an interesting comparison to the procedures discussed here. In the two-class setting, SC amounts to linear classification with feature weights obtained from soft thresholding of feature Z-scores. Consequently, HCT-soft can be viewed as a modification to SC, choosing thresholds by HCT rather than cross-validation. We made a simulation study contrasting the performance of SC with HCT-hard, HCT-soft, and HCT-clip in the Rare/Weak features model. We conducted 100 Monte-Carlo simulations, where we chose $p = 10000$, $k = 100$ (so $\epsilon = k/p = .01$), $n = 40$, and $\tau \in [1, 3]$. Over this range, the best classification error rate ranged from nearly 50%

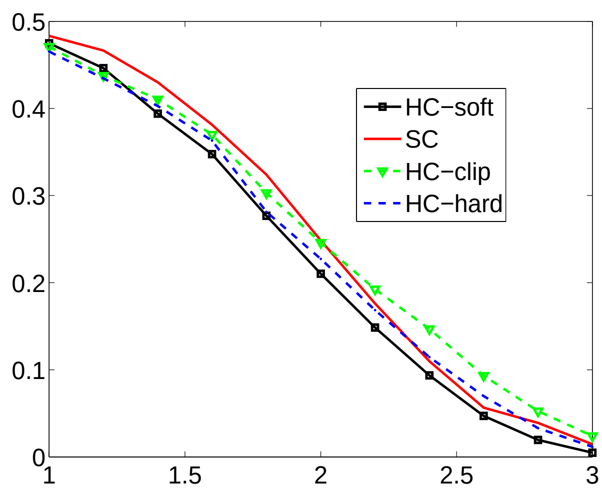

Fig. 5. Comparison of error rates by using Shrunken Centroids, threshold choice by cross validation, and linear classifiers using HCT-based threshold selection. Simulation assuming the RW model. Black: HCT-soft. Red: Shrunken Centroids. Green: HCT-clip. Blue: HCT-hard. x-axis displays τ.

– scarcely better than ignorant guessing – to less than 3%. Figure 5 shows the results. Apparently, HCT-soft and SC behave similarly – with HCT-soft consistently better (here SC is implemented with a threshold picked by 10-fold cross validations). However, HCT-soft and SC are not at all similar in computational cost at the training stage, as HCT-soft requires no cross-validation or tuning. The similarity of the two classifiers is, of course, explainable using discussions above. Cross-validation is 'trying' to estimate the ideal threshold, which the HCT functional also approximates. In Table 2, we tabulated the mean and standard deviation (SD) of HCT and cross-validated threshold selection (CVT). We see that CVT is on average larger than the HCT in this range of parameters. We also see that CVT has a significantly higher variance than the HC threshold; presumably this is why HCT-soft consistently outperforms SC. In fact, cross-validation is generally inconsistent in the fixed-n, large-p limit, while HCT is consistent in the RW model; hence the empirical phenomenon visible in these graphs should apply much more generally.

Alternative Classifier Using HC. HC can be used directly for classification [19], without reference to linear discrimination and feature selection. For comparison with the method proposed here, see [12].

Theoretical Optimality. In companion work [12], we develop a large-p, fixed-n asymptotic study and show rigorously that HCT yields asymptotically optimal error rate classifiers in the RW model.

Reproducible Research. All original figures and tables presented here are fully reproducible, consistent with the concept of Reproducible Research [15]. Software is available at [20].

ACKNOWLEDGMENTS. The authors would like to thank the Newton Institute for hospitality during the program **Statistical Challenges of High-Dimensional Data**, and particularly D.M. Titterington for his leading role in organizing this program. Peter Hall gave helpful comments on the behavior of cross-validation in the large p, small n case. DD was supported in part by NSF Grant DMS-0505303. JJ was supported in part by NSF Grant DMS-0505423 and DMS-0639980.

1. Abramovich F, Benjamini Y, Donoho D, Johnstone I (2006) Adapting to unknown sparsity by controlling the false discovery rate. Ann. Statist. 34, 584–653.
2. Abramovich F, Benjamini Y (1995) in Wavelets and Statistics, eds Antoniadis A, Oppenheim, G (Springer, NY), pp 5–14.
3. Alon U et al. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc. Natl. Acad. Sci. USA 96, 6745–6750.
4. Anderson, TW (2003) An Introduction to multivariate statistical analysis, 3rd ed. (Wiley, New York).
5. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. Roy. Statist. Soc. B 57, 289-300.
6. Bickel P, Levina E (2004) Some theory of Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. Bernoulli 10, 989–1010.
7. Breiman L (2001) Random forests. Mach. Learn 24, 5–32.
8. Burges C (1998) A tutorial on support vector machines for pattern recognition. Knowl. Discov. Data Min. 2, 121–167.
9. Dettling M (2004) BagBoosting for tumor classification with gene expression data. Bioinformatics 20, 3583–3593.
10. Dettling M, Bühlmann P (2003) Boosting for tumor classification with gene expression data. Bioinformatics 19, 1061–1069.
11. Donoho D, Jin J (2004) Higher criticism for detecting sparse heterogeneous mixtures. Ann. Statist. 32, 962-994.
12. Donoho D, Jin J (2008) Analysis of feature selection by Higher Criticism. Working manuscript.
13. Donoho D, Johnstone I (1994) Minimax risk over l_p-balls for l_q-error. Probab. Theory Related Fields 2, 277–303.
14. Donoho D, Johnstone I, Hoch JC, Stern AS (1992) Maximum entropy and the nearly black object. J. Royal Stat. Soc. B 54, 41-81.
15. Donoho D, Maleki A, Ur-Rahman I, Shahram M, Stodden V (2008) 15 years of reproducible research in computational harmonic analysis. Technical Report, Department of Statistics, Stanford.
16. Efron B, Tibshirani R, Storey J, Tusher, V (2001) Empirical Bayes analysis of a microarray experiment. J. Amer. Statist. Assoc. 99, 96-104.
17. Fan J, Fan Y (2008) High dimensional classification using features annealed independence rules. Ann. Statist., to appear.
18. Golub T et al. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286, 531–536.
19. Hall P, Pittelkow Y, Ghosh M (2008) Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes. J. Roy. Statist. Soc. B 70 158–173.
20. http://www-stat.stanford.edu/~donoho/software.
21. Hedenfalk I et al. (2001) Gene-expression profile in hereditary breast cancer. N. Engl. J. Med. 344, 539–448.
22. Ingster YI (1997) Some problems of hypothesis testing leading to infinitely divisible distribution. Math. Methods Statist. 6, 47–69.
23. Jin J (2003) Detecting and estimating sparse mixtures. Ph.D. Thesis, Department of Statistics, Stanford University.
24. Singh D et al. (2002) Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1, 203–209.
25. Tibshirani R, Hastie T, Narasimhan B, Chu G (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl. Acad. Sci. USA 99, 6567–6572.

Table 1. Error rates of Standard Classifiers on Standard Examples from Dettling [9]

Method	ALL/reg	Col/reg	Pro/reg	m-reg	R
Bagboo	4.08/.59	16.10/.52	7.53/0	.59	6
Boost	5.67/1	19.14/1	8.71/.18	1	7.5
RanFor	1.92/.02	14.86/.32	9.00/.22	.32	2
SVM	1.83/0	15.05/.35	7.88/.05	.35	3
DLDA	2.92/.28	12.86/0	14.18/1	1	7.5
KNN	3.83/.52	16.38/.56	10.59/.46	.56	5
PAM	3.55/.45	13.53/.11	8.87/.20	.45	4
HCT	2.86/.27	13.77/.14	9.47/.29	.29	1

Table 2. Comparison of HCT and CVT

τ	HCT(mean)	CVT(mean)	HCT(SD)	CVT(SD)
1.0	2.2863	3.8192	0.3746	1.9750
1.4	2.2599	3.3255	0.3401	1.7764
1.8	2.2925	3.0943	0.3400	1.3788
2.2	2.3660	2.6007	0.2921	0.8727
2.6	2.5149	2.5929	0.2644	0.5183
3.0	2.6090	2.9904	0.2698	0.5971