A tail strength measure for assessing the overall significance in a dataset

Jonathan Taylor[*]
Robert Tibshirani [†]

July 18, 2005

Abstract

We propose an overall measure of significance for a set of hypothesis tests. The *tail strength* is a simple function of the p-values computed for each of the tests. This measure is useful, for example, in assessing the overall univariate strength of a large set of features in microarray and other genomic and biomedical studies. It also has a simple relationship to the false discovery rate of the collection of tests. We derive the asymptotic distribution of the tail strength measure, and illustrate its use on a number of real datasets.

1 Introduction

Dave et al. (2004) published a study correlating the expression of $49,000$ genes from microarrays with patient survival in follicular lymphoma. The authors derived a multivariate Cox model for the data, and reported that it was highly predictive in an independent test set. Tibshirani (2005) re-analyzed this data, shedding considerable doubt on the reproducibility of the findings.

The left panel of Figure 1 shows the ordered Cox scores $T_{(k)}$ for each gene, plotted against the expected (null) order statistics $\mathbb{E}(T^*_{(k))})$, where the expectation is estimated by repeated permutations of the patient labels. We see that there is little deviation from the expected values. The right panel shows a similar plot for the leukemia data of Golub et al. (1999). This is a two class problem, so the scores $T_{(k)}$ are the ordered two-sample T-statistics. There are many more large values than we would expect to see by chance. Perhaps this is why the Golub dataset has become the most common testing ground for authors proposing new methods for microarray analysis.

[*] Department of Statistics, Stanford University, Stanford, CA 94305; jtaylor@stat.stanford.edu

[†] Department of Health Research & Policy and Department of Statistics, Stanford University, Stanford, CA 94305; tibs@stanford.edu

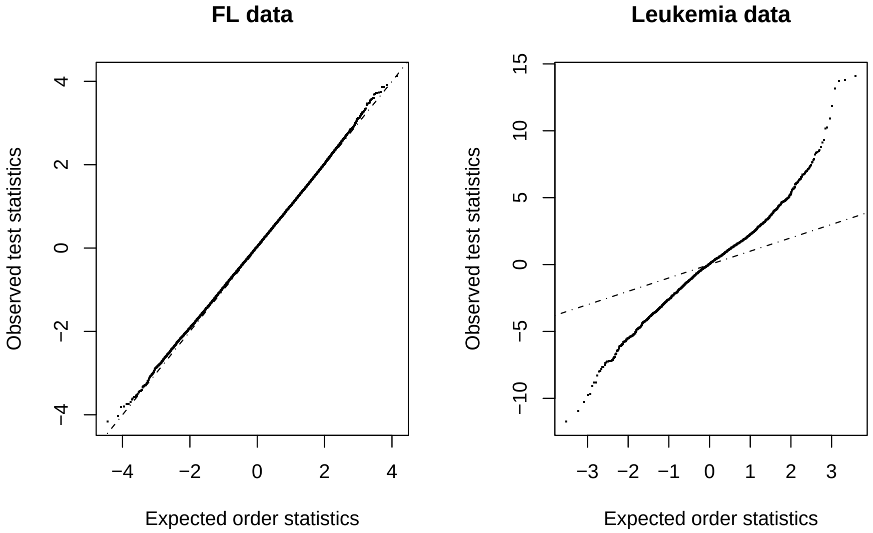

Figure 1: *Test statistics (one per gene) from the follicular lymphoma data (left) and leukemia data (right). Each plot shows the observed test statistics versus the expected order statistics under the null hypothesis.*

In the re-analysis of the Dave et al. (2004) data, it became clear that if there was predictive power in this dataset, it was very subtle. As seen in the left panel of Figure 1, the univariate effects of the genes seem to be very small. From this experience, we felt it would be useful to have a general quantitative measure of the univariate strength of a large set of predictors. Such a measure could be routinely reported, as an indication of the predictive strength in a dataset. Of course such a measure would not capture any multivariate or interactive effects that might be present.

In this paper we propose a measure of overall significance, called the "tail strength". We derive its asymptotic distribution and illustrate its use on a number of real datasets. We also relate our measure to the false discovery rate and the area under the ROC curve.

2 Tail strength

2.1 Definition

We first define our measure based on a set of p-values. Later, we give an equivalent form in terms of test statistics. We assume that we have null hypotheses H_{0i}, and associated p-values p_i $i = 1, 2, \ldots m$. Let the ordered p-values be $p_{(1)} \leq p_{(2)} \cdots \leq p_{(m)}$.

We define the *Tail Strength* as

$$\mathrm{TS}(p_1,\ldots,p_m)=\frac{1}{m}\sum_{k=1}^{m}\left(1-p_{(k)}\frac{m+1}{k}\right). \tag{1}$$

Now under the global null hypothesis, each p_k has a uniform distribution, so that the expected value of the kth smallest $p_{(k)}$ is $k/(m+1)$ and TS has expectation zero. The tail strength measures the deviation of each p-value from its expected value: $p_{(k)} < k/(m+1)$ causes $1-p_{(k)}\frac{m+1}{k}$ to be > 0. Thus large positive values of TS indicate evidence against the null hypothesis, that is, it indicates that there are more small p-values than we would expect by chance. Note also that the particular form of TS will give more weight to the lowest p-values, so that it is most sensitive to deviations in the tail.

For the FL and leukemia datasets, TS equals -0.027 and 0.655, respectively. Hence the FL p-values are slightly *larger* than we would expect under the uniform distribution. In contrast, the leukemia genes are highly significant.

Figure 2 shows the tail strength measure applied to some simulated microarray data. There are 1000 genes (features) and 20 samples; all measurements are standard $N(0,1)$, except for the first 100 genes in the second 10 samples, which were generated as $N(\Delta,1)$. The Figure shows tail strength divided by its standard error, from 100 realizations at each of 7 different values of Δ. We see that TS has the desired behavior: it is centered around zero, when the overall null hypothesis holds ($\Delta=0$) and then becomes more and more positive as Δ increases.

2.2 Tail strength for test statistics

There is an equivalent form for tail strength in terms of test-statistics $T_k, k = 1,\ldots,m$. Suppose we have a null distribution Prob_0 for these statistics, derived from a set of permutations or asymptotic theory. This yields a set of p-values

$$p_{(k)}=\mathrm{Prob}_0(|T|^*\geq|T_{(m-k)}|) \tag{2}$$

where $|T_{(1)}|\leq|T_{(2)}|\ldots\leq|T_{(m)}|$ are the test statistics ordered by absolute value. Then

$$\mathrm{TS}=\frac{1}{m}\sum_{k=1}^{m}\left(\frac{k-p_{(k)}\cdot(m+1)}{k}\right). \tag{3}$$

Each term is the proportion of test statistics that exceed the expected number, when testing at value $T_{(k)}$. Thus for example the value 0.655 for the leukemia data indicates that there are (on average) 65.5% more significant test statistics than we would expect by chance.

2.3 Relationship to FDR

The quantity TS is closely related to the False Discovery Rate (FDR) (Benjamini & Hochberg 1985, Efron et al. 2001, Storey 2002, Efron & Tibshirani 2002, Genovese & Wasserman 2002).

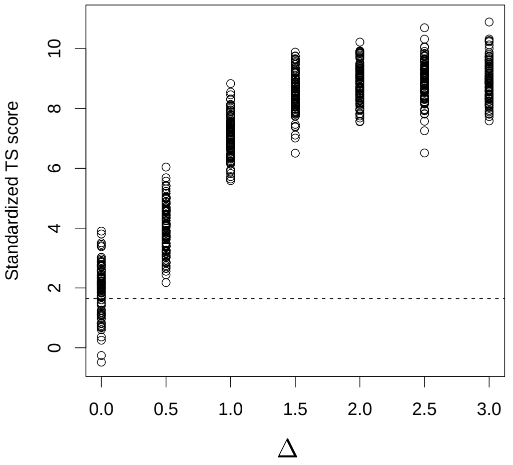

Figure 2: *Simulated microarray data: 1000 genes and 20 samples, and we wish to compare the first ten samples to the second ten. All measurements are standard Gaussian, except for the first 100 genes in samples 11-20, which have mean Δ. Shown are 100 realizations of tail strength divided by its standard error, at each value of Δ. A horizontal line is drawn at the upper 95% point 1.645.*

Table 1: *Possible Outcomes from m Hypothesis Tests*

	Accept	Reject	Total
Null True	U	V	m_0
Alternative True	Q	S	m_1
	W	R	m

We first review the FDR. Table 1 displays the various outcomes when testing m null hypotheses $H_{0i}, 1 \leq i \leq m$. The quantity V is the number of false positives (Type I errors), while R is the total number of hypotheses rejected, which depends on the testing procedure.

The false discovery rate (FDR) (Benjamini & Hochberg 1985) is defined the expected value of $V/R \cdot 1_{\{R>0\}}$. If the decision rule is a thresholding rule, then one can define the following plug-in estimate of FDR at p-value x (Storey (2002), Storey et al. (2004))

$$\widehat{\mathrm{FDR}}(x) = \frac{x}{\widehat{F}_m(x)} \cdot 1_{\{\widehat{F}_m(x)>0\}}, \qquad 0 \leq x \leq 1 \tag{4}$$

where

$$\widehat{F}_m(x) = \frac{\#\{p_i : p_i \leq x\}}{m} \tag{5}$$

the empirical CDF of the p-values $p_1, \ldots, p_m$.

There is a Bayesian model for this setting, that we will find useful in our later analysis. Given a prior *null* probability, π_0 and an alternative distribution F_1, the Bayesian model for observing m p-values is the following: for $1 \leq i \leq m$ independently

1. generate $H_{0,i} \sim \mathrm{Bernoulli}(\pi_0)$;
2. if $H_{0,i} = 0$, generate $p_i \sim \mathrm{Unif}(0,1)$, else generate $p_i \sim F_1$.

Under this model, it is easy to see that the p-values are unconditionally i.i.d. with distribution

$$F = \pi_0 \cdot \mathrm{Unif}(0,1) + (1-\pi_0) \cdot F_1.$$

Further, without any constraint on F_1, the parameter π_0 is obviously unidentifiable.

In Efron & Tibshirani (2002) and Storey (2002) it is shown that under this model

$$\mathbb{E}\left(\widehat{\mathrm{FDR}}(x) \middle| \widehat{F}(x) > 0\right) = \frac{\pi_0 \cdot x}{F(x)}.$$

For extensions to large samples, see Storey et al. (2004).

Finally we can derive the relationship between tail strength and FDR. Looking at the plug-in estimate (4), it is easy to see that

$$\mathrm{TS} = \sum_{k=1}^{m} \left(1 - p_{(k)} \cdot \frac{m+1}{k}\right) \simeq \sum_{k=1}^{m} \left(1 - p_{(k)} \cdot \frac{m}{k}\right) = \sum_{k=1}^{m} [1 - \widehat{\mathrm{FDR}}(p_k)]. \tag{6}$$

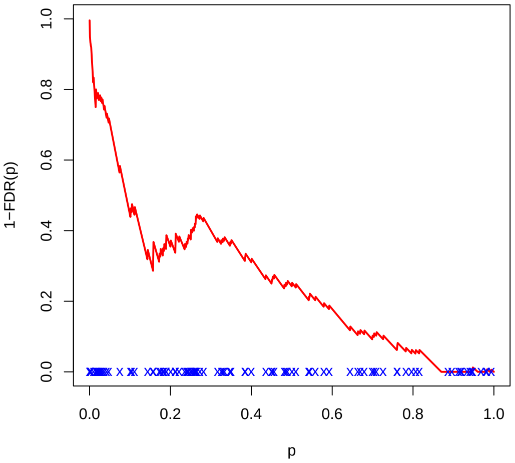

Figure 3: *A graphical description of tail strength. It measures a kind of area under the curve $1-FDR(x)$ (red), evaluating this function at the observed p-values p_k (blue).*

Figure 3 gives a graphical interpretation of the simple relationship (6): TS measures a kind of area under the curve $1-\widehat{\mathrm{FDR}}(x)$, evaluating this function at the observed p-values p_k. Hence the faster $1-\widehat{\mathrm{FDR}}(x)$ goes to one ($\widehat{\mathrm{FDR}}(x)$ drops to zero) as $x \downarrow 0$, the higher TS. Further, the tighter the p-values are bunched up near 0, the larger TS.

Another way of seeing that TS can be phrased directly in terms of the test statistics (as in (3)) comes from the fact that the expression $\widehat{\mathrm{FDR}}$ in (6) can be computed on the scale of the test statistics or the p-values. Therefore, TS is unchanged under any one-to-one transformation of the p-values, and is not tied to the choice of test statistic used to test each null hypothesis $H_{0i}, 1 \leq i \leq m$.

When the p-values are i.i.d. with distribution F, the following result, proven in Section 3, is therefore not surprising

$$\mathbb{E}(\mathrm{TS}) \overset{m \to \infty}{=} \mathbb{E}(1-\mathrm{FDR}(X)), \qquad X \sim F \tag{7}$$

where

$$\mathrm{FDR}(x) = \frac{x}{F(x)} \tag{8}$$

is the population FDR, with the unknown proportion of true null hypotheses π_0 set to 1. In other words, the tail strength statistic estimates the average amount by which the true false discovery rate function falls below its null value of 1, with the average computed with respect to the true distribution of p-values.

If F is stochastically dominated by $\mathrm{Unif}(0,1)$ then TS is asymptotically normal with variance

$$\mathrm{Var}(\mathrm{TS}) \overset{m\to\infty}{\simeq} \frac{C(F)}{m} \tag{9}$$

where $C(F) \leq 1$ if $F(x) \geq x$ for each x in $[0,1]$. We use the approximation

$$\mathrm{Var}(\mathrm{TS}) \approx 1/m \tag{10}$$

in all of our applications of tail strength.

Note that the quantity $m_1 = m - m_0$ measures how many non-null genes there are in the dataset. Various authors have studied this as a measure of univariate strength (cf. (Benjamini & Hochberg 2000, Storey et al. 2004)). However, this does not really measure *how different* the non-null p-values are from $\mathrm{Unif}(0,1)$. Further, in the Bayesian model described earlier, this parameter is not identifiable without some constraint on the alternatives. In contrast, tail strength is identifiable and measures *how far* the non-null p-values are from $\mathrm{Unif}(0,1)$.

3 Asymptotic properties of tail strength

In the false discovery rate (FDR) setting, previous work has shown that examination of the limiting behavior of (estimates of) FDR and local FDR is useful in understanding what the various techniques are doing in a *population* setting. In this section, we carry out a similar analysis for TS.

We can write

$$\begin{aligned}
\mathrm{TS} &= -\frac{1}{m}\sum_{k=1}^{m}\left(p_{(k)} - \frac{k}{m+1}\right)\frac{m+1}{k} \\
&\simeq -\frac{1}{m}\sum_{k=1}^{m}\left(p_{(k)} - \frac{k}{m}\right)\frac{m}{k} \\
&= -\frac{1}{m}\sum_{k=1}^{m}\sum_{j=1}^{k}\frac{m}{k}\left(p_{(j)} - p_{(j-1)} - \frac{1}{m}\right) \\
&= -\sum_{j=1}^{m}\left(p_{(j)} - p_{(j-1)} - \frac{1}{m}\right)\left(\sum_{k=j}^{m}\frac{1}{k}\right).
\end{aligned}$$

Under H_0, the spacings of order statistics are distributed as

$$s_j = p_{(j)} - p_{(j-1)} \sim \frac{\xi_j}{\sum_{i=1}^{m+1}\xi_i}$$

where $\xi_i \sim \mathrm{Exp}(1)$ are i.i.d. exponential random variables.

This suggests that TS should be asymptotically normally distributed, at least under H_0, because it is the sum of approximately independent random variables. In fact, TS is also asymptotically normally distributed when the p-values are identically distributed with distribution F, as in the Bayesian model of Storey (2002). We could alternatively assume that some fixed proportion π_0 of the p-values are i.i.d. Unif$(0,1)$, and the remaining are i.i.d. from some distribution F_1 such that

$$F = \pi_0 \cdot \mathrm{Unif}(0,1) + (1-\pi_0) \cdot F_1.$$

This is the mixture model used in the development of local FDR by Efron et al. (2001) and Efron & Tibshirani (2002). This assumption would not likely change the essence of our main result, only complicate the proofs.

We begin by expressing (1) in yet another way, in terms of quantile processes. Let

$$\widehat{Q}_m(q) = \widehat{F}_m^{-1}(q) = \inf\{x : \widehat{F}_m(x) > q\} \tag{11}$$

be the quantile process of the p-values and

$$Q(q) = F^{-1}(q) = \inf\{x : F(x) > q\} \tag{12}$$

be the population quantile function.

Given the definition of Q_m, it is not hard to see that

$$\widehat{Q}_m\left(\frac{k}{m}\right) = p_{(k)}, \qquad \forall\ 1 \leq k \leq m.$$

Using this fact, the expression (1) takes the form of a Riemann sum

$$\begin{aligned} \mathrm{TS}(p) &\simeq \frac{1}{m}\sum_{k=1}^{m}\left(1 - p_{(k)}\frac{m}{k}\right) \\ &= \frac{1}{m}\sum_{k=1}^{m}\frac{(k/m - Q(j/m)) + \left(Q(j/m) - \widehat{Q}_m(j/m)\right)}{k/m} \\ &\overset{m\to\infty}{\rightarrow} \int_0^1 1 - \frac{Q(x)}{x}\,dx + \int_0^1 \frac{Q(x)-\widehat{Q}_m(x)}{x}\,dx. \end{aligned} \tag{13}$$

Such an expression is simpler to analyze than (1), using some results from the theory of quantile processes.

If Q is Riemann integrable, the first expression converges to

$$\int_0^1 \left(1 - \frac{Q(x)}{x}\right)\,dx.$$

If, further, F has a density, then making the substitution $u = Q(x)$ we see that this expression is equal to (7).

The result we will use from the theory of quantile processes (Barrio 2004) is the following: under H_0

$$m^{1/2}\left(\widehat{Q}_m(x) - Q(x)\right)_{0\leq x\leq 1} \overset{D}{=} (B_x)_{0\leq x\leq 1} \tag{14}$$

where $(B(x))_{0\leq x\leq 1}$ is a standard Brownian bridge. That is, a continuous Gaussian process on $[0,1]$ with mean 0 and covariance function

$$\mathrm{Cov}(B(x),B(y)) = \min(x,y) - xy, \qquad 0 \leq x,y \leq 1. \tag{15}$$

Suppose the p-values are i.i.d. with distribution F, where F is twice-differentiable with strictly positive density f on $(0,1)$ then (Barrio 2004)

$$m^{1/2}\left(\widehat{Q}_m(x) - Q(x)\right)_{0\leq x\leq 1} \Rightarrow \left(\frac{B_x}{f(Q(x))}\right)_{0\leq x\leq 1}. \tag{16}$$

This suggests that

$$m^{1/2}(\mathrm{TS} - \mathbb{E}(\mathrm{TS})) \Rightarrow \int_0^1 \frac{B_x}{xf(Q(x))}\,dx \overset{D}{=} \int_0^1 \frac{B_{F(x)}}{F(x)}\,dx.$$

A straightforward application of Theorem 1 of (Shorack 1972), combined with the comments above suffices to prove the following result.

Theorem 3.1 *Under the Bayesian model of Storey (2002), suppose that $F(x) \geq x$. Then, if F has density f, as $m\to\infty$,* TS *is asymptotically normally distributed with mean*

$$\mathbb{E}(\mathrm{TS}) \overset{m\to\infty}{=} \mathbb{E}(1 - FDR(X)), \qquad X \sim F$$

and variance

$$Var(m^{1/2}(\mathrm{TS} - \mathbb{E}(\mathrm{TS})) \overset{m\to\infty}{\to} \int_0^1\int_0^1 \left(\frac{\min(x,y)}{xy} - 1\right)\,dQ(x)\,dQ(y).$$

Remark: Actually, F need not even have a density, for the central limit to hold above, though the expected value will be changed slightly. If F has density f, then under the hypothesis $F(x) \geq x$

$$\begin{aligned}\int_0^1\int_0^1 \left(\frac{\min(F(x),F(y))}{F(x)F(y)} - 1\right)\,dx\,dy &= 2\cdot\int_0^1\int_y^1 \left(\frac{1}{F(x)} - 1\right)\,dx\,dy\\ &\leq 2\cdot\int_0^1\int_y^1\left(\frac{1}{x} - 1\right)\,dx\,dy\\ &= 1\end{aligned}$$

so the variance under H_0 is an upper bound.

4 Relationship of TS to area under the ROC curve

In the diagnostic testing literature (cf. (Hanley & McNeil 1982, Pepe 2003)), the ROC curve is used to discriminate between two samples. Such a curve can also be constructed to compare a sample of test statistics to a given null distribution. A commonly used summary of the ROC curve is the area under the ROC curve. In the two-sample setting, the area is essentially equivalent to the Mann-Whitney test statistic (Hanley & McNeil 1982). This measure places equal weight on departures from Unif$(0,1)$ without focusing on the most interesting region, the *tail* of the test statistics. One solution is to only look at the area under the ROC curve up to some false positive level t_0 (Pepe 2003), but the choice of t_0 is somewhat arbitrary. Here we show that the tail strength measure is related to a weighted area under such an ROC curve, weighted to accentuate the tail of the test statistics.

It is well known (Hanley & McNeil 1982) that for two independent samples $\{p_1, \dots, p_{m_0}\} \overset{IID}{\sim}$ Unif$(0,1)$ and $\{q_1, \dots, q_{m_1}\} \overset{IID}{\sim} F$ that the expected area under the empirical ROC curve

$$\widehat{ROC} = \{(\widehat{F}_{m_0}(x), \widehat{G}_{m_1}(x)), x \in [0,1]\}$$

is

$$\text{Prob}\{X \le Y\}, \qquad X \sim \text{Unif}(0,1), Y \sim F.$$

The measure TS is also closely related to the area under the ROC curve (Pepe 2003, Hanley & McNeil 1982). Let

$$ROC = \{(F(x), x), x \in [0,1]\}$$

be the population ROC curve reflected along the line $y = x$. Suppose that $X \sim \text{Unif}(0,1)$ and $Y \sim F$, then

$$\Pr\{X \le Y\} = \int_0^1 \Pr(X \le y \big| Y = y) f(y)\, dy = \int_0^1 y f(y)\, dy$$

and this quantity is $\frac{1}{2}$ if $F = \text{Unif}(0,1)$. This suggests that the area under the (ROC) curve

$$AUC - \frac{1}{2} = \int_0^1 (x - F(x))\, f(x) dx = \int_0^1 \left(F^{-1}(x) - x\right)\, dx \tag{17}$$

is a measure of departure from uniformity. It is positive whenever $F(x) \le x$, or whenever the p-values are stochastically dominated by Unif$(0,1)$. This quantity places equal weight on the differences for all values of x with no focus on the tail. One way to adjust it is to insert a weight into the expression (17)

$$AUC - \frac{1}{2} = \int_0^1 (x - F(x)) w(x)\, f(x) dx \tag{18}$$

The choice $w(x) = x^{-1}$ corresponds to TS, in the asymptotic setting. In finite samples, the integral above is of course replaced by a Riemann sum.

The partial AUC proposed by Pepe (2003) also attempts to accentuate the tail

$$pAUC(p_0) = \int_0^{p_0} (F^{-1}(x) - x)\, dx.$$

This is of course equivalent to choosing a weight

$$w(x) = 1_{[0,p_0]}(x)$$

while ((17)). Setting $w(x) = x^{-1}$, which puts more weight on the tail, yields TS.

5 Real data examples

Figure 4 shows the tail strength measure and asymptotic 90% confidence intervals, applied to 12 different datasets. The datasets are summarized in Table 2. The first 9 datasets are from microarray studies, and all report positive findings. Most of these are described Dettling (2004), where some comparative analyses are also performed.

The remaining datasets are from neuroimaging studies. The datasets *aud-over* and *aud-sent* are from an auditory fMRI study (Taylor & Worsley 2005) with *aud-over* being overall activation, and *aud-sent* a measure of hemodynamic delay (Liao et al. n.d.) in response to different sentences. The dataset *dtiTS* comes from a DTI (Diffusion Tensor Imaging) dataset, studying pediatric differences in white matter in dyslexic and control cases (Deutsch et al. 2005), the p-values reflect local differences in direction of white matter fiber tracts and were studied in Schwartzmann et al. (2005).

All of the datasets (except for FL) show significant (non-zero) tail strengths of various degrees. For the subset of classification problems among these studies, Table 3 compares the estimated tail strength with the misclassification rate from the nearest shrunken centroid classifier (Tibshirani et al. 2001) [Results from other classifiers, given in (Dettling 2004), are quite similar]. The error rates were computed by repeated $(2/3, 1/3)$ train-test splits of the data, except for the `skin` data which uses 14 fold cross-validation.

There is one interesting (qualitative) discrepancy in Table 3: the multi-class `brain` dataset shows very different behavior in tail strength and misclassification rate. The tail strength is high— 0.82 , but the misclassification rate seems poor (23.5%). The test statistic for each gene is an F-statistic— the ratio of between-class to within-class variance. Figure 5 shows the ordered test statistics versus their expected values under the null hypothesis. There is clearly more variation that we would expect by chance.

There are some possible explanations for the seeming discrepancy between tail strength and classification rate in the `brain` example. First note that with five classes, the base error rate is 80%, so that the value 23.5% is actually a substantial reduction in this rate. In addition, there only 42 cases in this dataset, so that the training set on which the classifiers were trained had only 28 cases on the average. For the five classes, the class-wise error rates were $(15, 6, 6, 20, 57)\%$. We computed the tail strengths for each class versus the rest (based on a two-sample t-statistic): they were $(0.39, 0.53, 0.67, 0.54, 0.32)$. Hence class 5 has both a high error rate and a lower tail strength. It seems

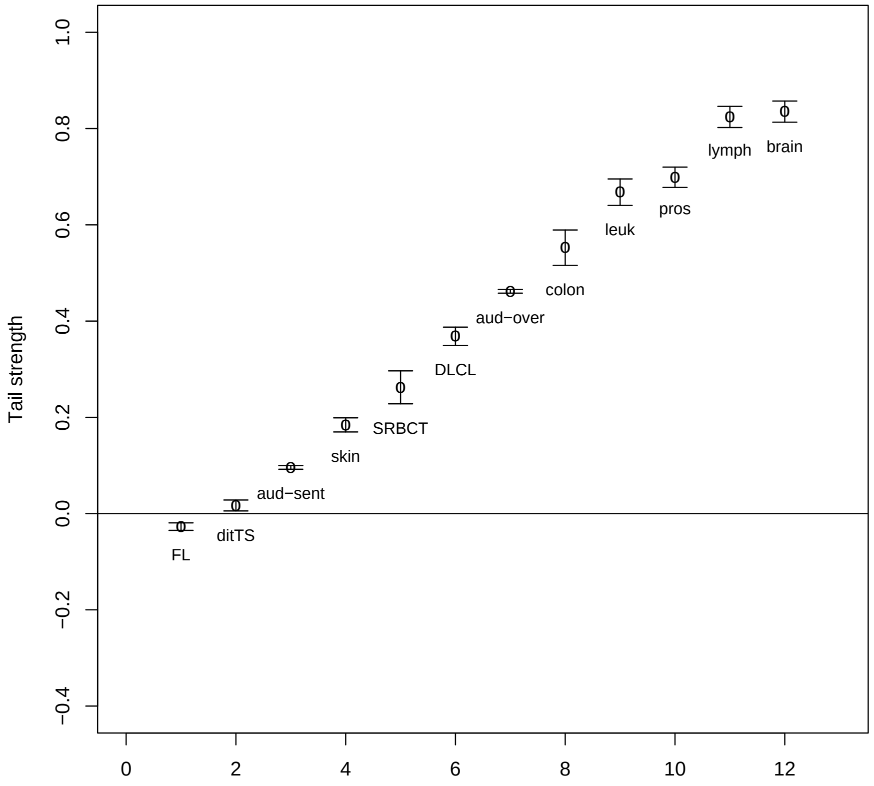

Figure 4: *Tail strength measure computed on some real datasets. Shown are the* TS *measures along with 90% confidence intervals.*

Name	Description	# Samples	# Features	Source
Follicular lymphoma	Microarray, Survival	93	44,928	Dave et al. (2004)
Skin cancer	Microarray, Two classes	58	12625	Rieger et al. (2004)
Diffuse large cell lymphoma	Microarray, Survival	240	7399	Rosenwald et al. (2002)
Small round blue cell tumors	Microarray, Four classes	63	2308	Khan et al. (2001)
Colon cancer	Microarray, Two classes	62	2000	Alon et al. (1999)
Leukemia	Microarray, Two classes	72	3571	Golub et al. (1999)
Prostate cancer	Microarray, Two classes	102	6033	Singh et al. (2002)
Brain cancer	Microarray, Five classes	22	5597	Pomeroy et al. (2002)
Lymphoma	Microarray, Two classes	62	4026	Alizadeh et al. (2000)
aud-over	FMRI	15	187782	Taylor & Worsley (2005)
aud-sent	FMRI	15	187762	Taylor & Worsley (2005)
dtiTS	Diffusion tensor imaging	12	20931	Schwartzmann et al. (2005)

Table 2: *Summary of datasets for Figure 4*

.

Dataset	Tail strength	% Misclassification rate
Lymphoma	0.82	1.7
Brain	0.82	23.5
SRBCT	0.77	2.4
Prostate	0.70	8.9
Leukemia	0.68	3.6
Colon	0.55	13.5
Skin	0.18	20.1

Table 3: *Tail strengths and misclassification rates (test set or cross-validated), for the classification problems in Table 2. Classification was done using nearest shrunken centroids.*

that the overall tail strength, based on the F-statistic for all five classes, fails to capture the difficulty in predicting class five.

6 Discussion

The tail strength measure seems to be potentially useful for assessing the overall significance of a set of hypothesis tests. For example, it gives a quantitative idea of the overall univariate association between a large set features, such as the genes in a microarray study, and an outcome of interest. We suggest that the tail strength could be routinely reported in such studies, to give the reader a crude idea of the significance in a complex dataset.

In Statistics, there is of course a long history and a substantial literature in the area of multiple hypothesis testing With the flury of applications in genomics, there has been a resurgence of interest in this area: see e.g. Dudoit et al. (2000) for a summary. Our work has a close relationship to the false

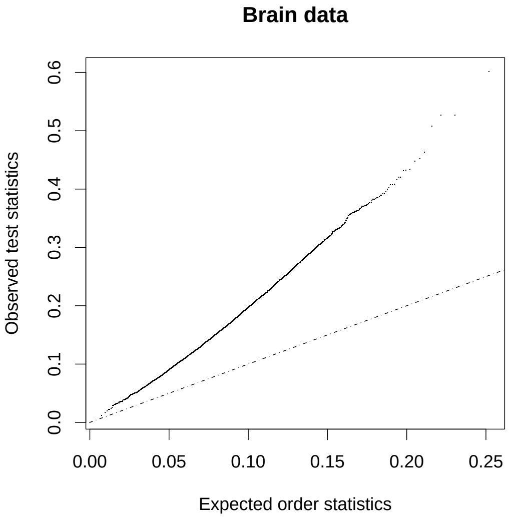

Figure 5: *Brain data: ordered statistics (F-statistics) versus their expected values under the null hypothesis.*

discovery rate approach to multiple testing, as we have shown in Section 2.3. There is recent work of **?** (Section 5), in which quantities similar to tail strength are considered, based on a local version of the false discovery rate.

Another concept that seems connected to tail strength is the *higher criticism* of Donoho & Jin (2004), generalizing an idea introduced by Tukey in 1976. They define

$$\mathrm{HC}_m = \max_{\{1\leq i\leq \alpha_0\cdot m\}}\sqrt{m}(i/m - p_{(i)})/\sqrt{p_{(i)}(1-p_{(i)})} \tag{19}$$

for some $\alpha_0 > 0$. This statistic is designed as an overall summary of the p-values, and they prove that is optimal for detecting certain sparse patterns of p-values. They also show that the asymptotic α percentile for HC_m is of size $\sqrt{\log\log m}$. We attempted some numerical comparisons of HC_m with tail strength on the datasets in this paper, but were not successful. The presence of some very small p-values made the denominator very small and caused the statistic to blow up. In addition, it wasn't clear how to choose α_0 and the significance cutpoint in finite samples. We leave this comparison for future study.

In summary, the tail strength measure proposed here is simple to compute, with no parameters that require adjustment. It must be stressed, however, that it does not measure all of the interesting structure that might be present in a dataset. When applied to univariate association measures, it does not capture interactions or multivariate effects that might exist.

Acknowledgments: We would like to thank Brad Efron for helpful comments. Tibshirani was partially supported by National Science Foundation Grant DMS-9971405 and National Institutes of Health Contract N01-HV-28183.

References

Alizadeh, A., Eisen, M., Davis, R. E., Ma, C., Lossos, I., Rosenwal, A., Boldrick, J., Sabet, H., Tran, T., Yu, X., J., P., Marti, G., Moore, T., Hudsom, J., Lu, L., Lewis, D., Tibshirani, R., Sherlock, G., Chan, W., Greiner, T., Weisenburger, D., Armitage, K., Levy, R. Wilson, W., Greve, M., Byrd, J., Botstein, D., Brown, P. & Staudt, L. (2000), 'Identification of molecularly and clinically distinct substypes of diffuse large b cell lymphoma by gene expression profiling', *Nature* **403**, 503–511.

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D. & Levine, A. (1999), 'Broad patterms of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays', *Proc. Nat. Acad. Sciences* **96**, 6745–6750.

Barrio, E. (2004), Empirical and quantile processes in the asymptotic theory of goodness of fit tests. Available online: $http$: $//www.eio.uva.es/ems/Goodness_of_fit - Laredo_2004.pdf$.

Benjamini, Y. & Hochberg, Y. (1985), 'Controlling the false discovery rate: a practical and powerful approach to multiple testing', *J. Royal. Stat. Soc. B.* **85**, 289–300.

Benjamini, Y. & Hochberg, Y. (2000), 'On the adaptive control of the false discovery fate in multiple testing with independent statistics', *J. Educ. Behav. Stat.* **25**(1), 60–83.

Dave, S. S., Wright, G., Tan, B., Rosenwald, A., Gascoyne, R. D., Chan, W. C., Fisher, R. I., Braziel, R. M., Rimsza, L. M., Grogan, T. M., Miller, T. P., LeBlanc, M., Greiner, T. C., Weisenburger, D. D., Lynch, J. C., Vose, J., Armitage, J. O., Smeland, E. B., , Kvaloy, S., , Holte, H., , Delabie, J., , Connors, J. M., Lansdorp, P. M., , Ouyang, Q., Lister, T. A., Davies, A. J., Norton, A. J., Muller-Hermelink, H. K., Ott, G., Campo, E., Montserrat, E., Wilson, W. H., , Jaffe, E. S., Simon, R., Yang, L., Goldschmidt, J. P. M. H. Z. M. N., Chiorazzi, M. & Staudt, L. M. (2004), 'Prediction of survival in follicular lymphoma based on molecular features of tumor-infiltrating immune cells', *The New England Journal of Medicine* **351**, 2159–2169.

Dettling, M. (2004), 'Bagboosting for tumor classification with gene expression data', *Bioinfoamatics* pp. 3583–3593.

Deutsch, G. K., Dougherty, R. F., Bammer, R., Siok, W. T., Gabrieli, J. D. & Wandell, B. (2005), 'Correlations between white matter microstructure and reading performance in children', *Cortex* **41**, 354–363.

Donoho, D. & Jin, J. (2004), ‘Higher criticism for detecting sparse heterogeneous mixtures’, *Annals of Statistics* (23), 962–994.

Dudoit, S., Yang, Y., Callow, M. & Speed, T. (2000), Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. Unpublished, available at http://www.stat.berkeley.edu/users/sandrine.

Efron, B. & Tibshirani, R. (2002), ‘Microarrays, empirical bayes methods, and false discovery rates’, *Gen. Epi.* .

Efron, B., Tibshirani, R., Storey, J. & Tusher, V. (2001), ‘Empirical bayes analysis of a microarray experiment’, *J. Amer. Statist. Assoc* pp. 1151–1160.

Genovese, C. & Wasserman, L. (2002), ‘Operating characteristics and extensions of the FDR procedure’, *J. Roy. Stat. Soc., Ser. B* **64**, 499–517.

Golub, T., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. & Lander, E. S. (1999), ‘Molecular classification of cancer: class discovery and class prediction by gene expression monitoring’, *Science* **286**, 531–536.

Hanley, J. A. & McNeil, B. J. (1982), ‘The meaning and use of the area under a receiver operating characteristic (ROC) curve’, *Radiology* **143**, 29–36.

Khan, J., Wei, J. S., Ringnér, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C. & Meltzer, P. S. (2001), ‘Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks’, *Nature Medicine* **7**, 673–679.

Liao, C., Worsley, K., Poline, J.-B., Aston, J., Duncan, G. & Evans, A. (n.d.), ‘Estimating the delay of the response in fMRI data’, *NeuroImage* **16**, 593–606.

Pepe, M. S. (2003), ‘Partial auc estimation and regression’, *Biometrics* **59**(3), 614–623.
*http://www.blackwell-synergy.com/doi/abs/10.1111/1541-0420.00071

Pomeroy, S., Tamayo, P., Gaasenbeek, M., Sturla, L., Angelo, M., McLaughlin, M., Kim, J., Goumnerova, L., Black, P., Lau, C., Allen, J., Zagzag, D., Olson, J., Curran, T., Wetmore, C., Biegel, J., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D., Mesirov, J., Lander, E. & Golub, T. (2002), ‘Prediction of central nervous system embryonal tumour outcome based on gene expression’, *Nature* **5**, 436–42.

Rieger, K., Hong, W., Tusher, V., Tang, J., Tibshirani, R. & Chu, G. (2004), 'Toxicity from radiation therapy associated with abnormal transcriptional responses to dna damage', *Proceedings of the National Academy of Sciences* **101**, 6634–6640.

Rosenwald, A., Wright, G., Chan, W. C., Connors, J. M., Campo, E., Fisher, R. I., Gascoyne, R. D., Muller-Hermelink, H. K., Smeland, E. B. & Staudt, L. M. (2002), 'The use of molecular profiling to predict survival after chemotherapy for diffuse large b-cell lymphoma', *The New England Journal of Medicine* **346**, 1937–1947.

Schwartzmann, A., Dougherty, R. & Taylor, J. (2005), 'Cross-subject comparison of principal diffusion direction maps', *Magnetic Resonance in Medicine* **53**, 1423–1431.

Shorack, G. R. (1972), 'Functions of order statistics', *Ann. Math. Statist.* **43**, 412–427.

Singh, D., Febbo, P., Ross, K., Jackson, D., Manola, J., Ladd, C., Tamayo, P., Renshaw, A., D'Amico, A., Richie, J., Lander, E., Loda, M., Kantoff, P., Golub, T. & Sellers, W. (2002), 'Gene expression correlates of clinical prostate cancer behavior', *Cancer cell* **1**, 203–9.

Storey, J. D. (2002), 'A direct approach to false discovery rates', *J. Roy. Stat. Soc., Ser. B* **64**, 479–498.

Storey, J. D., Taylor, J. E. & Siegmund, D. O. (2004), 'Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: A unified approach.', *J. Roy. Stat. Soc., Ser. B* **66**, 187–205.

Taylor, J. & Worsley, K. (2005), 'Analysis of hemodynamic delay in the FIAC data', Presented at HBM2005, Toronto. 2005.

Tibshirani, R. (2005), 'Immune signatures in follicular lymphoma', *The New England Journal of Medicine* **352**, 1496–1497.

Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. (2001), 'Diagnosis of multiple cancer types by shrunken centroids of gene expression', *Proc. Natl. Acad. Sci.* **99**, 6567–6572.