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

JONATHAN TAYLOR
Department of Statistics, Stanford University, Stanford, CA 94305, USA
jtaylor@stat.stanford.edu

ROBERT TIBSHIRANI*
Department of Health Research and Policy and Department of Statistics, Stanford University, Stanford, CA 94305, USA
tibs@stanford.edu

SUMMARY
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.

Keywords: Multiple testing; \( p \)-value.

1. INTRODUCTION
Dave et al. (2004) published a study correlating the expression of 44,928 clones from microarrays with patient survival in follicular lymphoma (FL). The authors derived a multivariate Cox model for the data, and reported that it was highly predictive in an independent test set. Tibshirani (2005) reanalyzed 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 (see, e.g. Kalbfleisch and Prentice, 1980). These are the partial likelihood score statistics, and are plotted against the expected (null) order statistics \( E(T^*(k)) \), 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 problem compares two disease classes, 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.

In the reanalysis 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, very few genes seem to exhibit

*To whom correspondence should be addressed.
univariate effects. 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, that is, the marginal effects. 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 the overall statistical significance in testing the global hypothesis of no gene effects. We call it 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 (FDR) 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 \( p_i, i = 1, 2, \ldots, m \). The global null hypothesis states that \( H_{0i} \) holds for all \( i \), and we assume that under this hypothesis, the \( p_i \) are i.i.d. \( U[0,1] \) random variables.

Let the ordered \( p \)-values be \( p_1 \leq p_2 \leq \cdots \leq p_m \). We define the tail strength as

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

Now under the global null hypothesis, each \( p_k \) has a uniform distribution, so that the expected value of the \( k \)th 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 global 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. In contrast, for example, the sum of successive differences in \( p \)-values would not have this property.

There is a Bayesian model for this setting that we will find useful in our later analysis. Given a prior ‘null’ probability, \( \pi_0 \), 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 $H_{0,i} \sim \text{Bernoulli}(\pi_0)$;
2. if $H_{0,i} = 0$, generate $p_i \sim \text{Unif}(0, 1)$, 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

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

Further, without any constraint on $F_1$, the parameter $\pi_0$ is obviously unidentifiable.

The most common application of tail strength is likely to be in assessing the univariate (marginal) effect of a set of predictors. Suppose we have predictors $x_{ij}, j = 1, 2, \ldots, m$, and response variable $Y_i$, for observations $i = 1, 2, \ldots, N$. Letting $y = (y_1, y_2, \ldots, y_N)$ and $x_j = (x_{1j}, \ldots, x_{Nj})$, we form a test statistic for each predictor:

$$T_j = f(x_j, y).$$  \hspace{1cm} (2.2)

Thus, $T_j$ measures the univariate effect of the $j$th predictor on the response. The two-sample $t$-statistic is a simple example.

There is an equivalent form of tail strength, expressed in terms of the $T_j$. Suppose we have a null distribution $\text{Prob}_0$ for these statistics, derived from a set of permutations or asymptotic theory. This yields a set of $p$-values

$$p_{(k)} = \text{Prob}_0(|T_k| \geq |T_{(m-k)}|),$$  \hspace{1cm} (2.3)

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

$$\text{TS} = \frac{1}{m} \sum_{k=1}^{m} \left( \frac{k - p_{(k)}(m + 1)}{k} \right).$$  \hspace{1cm} (2.4)

Each term is the proportion of test statistics that exceeds the expected number, when testing at value $T_{(k)}$.

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. 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.

The left panel of 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 plot shows tail strength divided by its standard error $1/\sqrt{1000}$, from 100 realizations at each of the seven different values of $\Delta$. 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 $\Delta$ increases.

In the right panel, the genes are not independent but have pairwise correlation 0.5 before $\Delta$ is added. Hence, the resulting $p$-values are exchangeable (but not independent) under the null hypothesis. We see that the expectation of tail strength behaves as in the independent case; however, its variance seems to be inflated. We address this issue in Section 3.

## 2.2 Relationship to FDR

The quantity TS is closely related to the FDR (Benjamini and Hochberg, 1995; Efron et al., 2001; Storey, 2002; Efron and Tibshirani, 2002; Genovese and Wasserman, 2002).

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.
Fig. 2. Simulated microarray data: 1000 genes and 20 samples, and we wish to compare the first 10 samples to the second 10. In the left panel, all measurements were generated as standard Gaussian, except for the first 100 genes in samples 11–20, which have mean $\Delta$. Shown are 100 realizations of tail strength divided by its standard error $1/\sqrt{100}$, at each value of $\Delta$. A horizontal line is drawn at the upper 95% point 1.645. In the right panel, the genes were not independent but were generated with pairwise correlation 0.5 before $\Delta$ was added. Hence, the resulting $p$-values are exchangeable (but not independent) under the null hypothesis.

Table 1. Possible outcomes from $m$ hypothesis tests

<table>
<thead>
<tr>
<th></th>
<th>Accept</th>
<th>Reject</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null true</td>
<td>$U$</td>
<td>$V$</td>
<td>$m_0$</td>
</tr>
<tr>
<td>Alternative true</td>
<td>$Q$</td>
<td>$S$</td>
<td>$m_1$</td>
</tr>
<tr>
<td></td>
<td>$W$</td>
<td>$R$</td>
<td>$m$</td>
</tr>
</tbody>
</table>

FDR (Benjamini and Hochberg, 1995) is defined as the expected value of $(V/R) \times 1_{(R>0)}$. 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)

$$\hat{\text{FDR}}(x) = \frac{x}{\hat{F}_m(x)} \times 1_{\{\hat{F}_m(x) > 0\}}, \quad 0 \leq x \leq 1,$$

where

$$\hat{F}_m(x) = \frac{\#\{p_i : p_i \leq x\}}{m},$$

the empirical cumulative distribution function of the $p$-values $p_1, \ldots, p_m$.

In Efron and Tibshirani (2002) and Storey (2002), it is shown that under the Bayesian model of section 2,

$$E(\hat{\text{FDR}}(x) | \hat{F}(x) > 0) = \frac{\pi_0 x}{F(x)}.$$

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 (2.5), it is easy to see that

$$\text{TS} = \frac{1}{m} \sum_{k=1}^{m} \left(1 - p(k) \frac{m+1}{k}\right) \simeq \frac{1}{m} \sum_{k=1}^{m} \left(1 - p(k) \frac{m}{k}\right) = \frac{1}{m} \sum_{k=1}^{m} [1 - \hat{\text{FDR}}(p_k)].$$
Tail strength measure for assessing univariate significance

Figure 3 gives a graphical interpretation of the simple relationship (2.7): TS measures a weighted area under the curve $1 - \hat{\text{FDR}}(x)$, evaluating this function at the observed $p$-values $p_k$ ($x$ values).

Another way of seeing that TS can be phrased directly in terms of the test statistics (as in (2.4)) comes from the fact that the expression $\hat{\text{FDR}}$ in (2.7) 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 $H_0$, $1 \leq i \leq m$.

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

$$\mathbb{E}(\text{TS}) \xrightarrow{m \to \infty} \mathbb{E}(1 - \text{FDR}_{\infty}(X)), \quad X \sim F,$$

(2.8)

where

$$\text{FDR}_{\infty}(x) = \frac{x}{F(x)}$$

(2.9)

is the (asymptotic) population FDR, with the unknown proportion of true null hypotheses $\pi_0$ set to one. In other words, the tail strength statistic estimates the average amount by which the true FDR function falls below its null value of one, with the average computed with respect to the true distribution of $p$-values.

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

$$\text{Var}(\text{TS}) \xrightarrow{m \to \infty} \frac{C(F)}{m},$$

(2.10)

where $C(F) \leq 1$ if $F(x) \geq x$ for each $x$ in [0, 1]. Hence, we have the asymptotic approximation

$$\text{Var}(\text{TS}) \approx 1/m.$$  

(2.11)

In Section 3, we examine the accuracy of this approximation.
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 and Hochberg, 2000; Storey et al., 2004). However, this does not really measure how different the non-null $p$-values are from 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 Unif$(0, 1)$.

The asymptotic behavior of tail strength is summarized in the following theorem.

**Theorem 2.1** 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

$$
\mathbb{E}(TS) \xrightarrow{m \to \infty} \mathbb{E}(1 - F\text{DR}(X)), \quad X \sim F,
$$

and variance

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

The proof appears in the Appendix.

### 2.3 Relationship of TS to area under the ROC curve

In the diagnostic testing literature (cf. Hanley and 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 and 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 and McNeil, 1982) that for two independent samples $\{p_1, \ldots, p_{m_0}\} \sim \text{IID Unif}(0, 1)$ and $\{q_1, \ldots, q_{m_1}\} \sim F$ the expected area under the empirical ROC curve

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

is

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

The measure $TS$ is also closely related to the area under the ROC curve (Pepe, 2003; Hanley and 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

$$
\text{Pr}\{X \leq Y\} = \int_0^1 \text{Pr}(X \leq y|Y = y) f(y) dy = \int_0^1 y f(y) dy
$$
and this quantity is $\frac{1}{2}$ if $F = \text{Unif}(0, 1)$. This suggests that the area under the (ROC) curve

$$\text{AUC} - \frac{1}{2} = \int_0^1 (x - F(x)) f(x) dx = \int_0^1 (F^{-1}(x) - x) dx$$  \hspace{1cm} (2.12)$$

is a measure of departure from uniformity. It is positive whenever $F(x) \leq x$ or whenever the $p$-values are stochastically dominated by $\text{Unif}(0, 1)$. 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 (2.12)

$$\text{AUC} - \frac{1}{2} = \int_0^1 (x - F(x))w(x)f(x) dx.$$  \hspace{1cm} (2.13)$$
The choice $w(x) = x^{-1}$ corresponds to TS in the asymptotic setting. In finite samples, the above integral is of course replaced by a Riemann sum.

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

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

Though the axes are reversed, this is equivalent to choosing a weight $w(x) = 1_{[0,p_0]}(x)$ while (2.12). Setting $w(x) = x^{-1}$, which puts more weight on the tail, yields TS.

### 3. Estimates of Variance

To use the tail strength measure in practice, we need a reasonably accurate estimate of its variance. Formula (2.11) is very simple, but assumes both independence of the genes and the global null hypothesis.

To assess the accuracy of formula (2.11), we did a simulation experiment. To ensure that the correlation structure was realistic, we used the gene expression data from Rieger et al. (2004), consisting of 12,625 genes and 58 samples in two classes. We constructed three scenarios: in the null scenario, datasets were created by permuting the sample labels (leaving the expression data intact); in the first non-null scenario, we first permuted the sample labels and then added 2000 units to the first 500 genes (2000 was about the largest average difference in group means in the actual data), and in the second non-null scenario, we added a random amount $u_i \sim N(0, 200^2)$ to 2000 genes $i$ chosen at random.

We carried out 100 realizations of this experiment, and the results are shown in Table 2. The quantity 'sd' is the actual standard deviation of tail strength over the 100 realizations; $\hat{\text{sd}}_0 = 1/\sqrt{12\,625}$, the asymptotic standard error. The third column $\hat{\text{sd}}_{\text{perm}}$ is a different estimate, one that starts with the permutation values used in the original computation of tail strength. We compute tail strength from successive blocks of 20 permutations, and then compute the standard deviation of the resulting tail strength values. This estimate explicitly assumes that the null hypothesis is true. We see that sd is much too small, in general, because of the lack of independence of the genes. On the other hand, the permutation-based estimate $\hat{\text{sd}}_{\text{perm}}$ is reasonably accurate under both the null and non-null scenarios. As we might expect, it is somewhat conservative (too large) under the non-null setup, but this is acceptable in practice. Alternatively, one could use a (non-null) bootstrap process to estimate the standard error, but this would require a great deal more computation. This might be an impediment against routine use of tail strength measure, so we use the permutation-based estimate $\hat{\text{sd}}_{\text{perm}}$ in the real data examples in this paper.
Table 2. Results of a simulation experiment to assess estimates' standard error of the tail strength measure. 'sd' is the actual standard deviation of tail strength over the 100 realizations; $\hat{sd}_0 = 1/\sqrt{12} = 0.527$, the asymptotic standard error; $\hat{sd}_{\text{perm}}$ is obtained by computing tail strength from successive blocks of 20 permutations, and then computing their standard deviation.

<table>
<thead>
<tr>
<th>Mean of TS</th>
<th>sd</th>
<th>$\hat{sd}_0$</th>
<th>Mean (sd) of $\hat{sd}_{\text{perm}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null setup</td>
<td>0.007</td>
<td>0.055</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.060 (0.010)</td>
</tr>
<tr>
<td>Non-null setup #1</td>
<td>0.123</td>
<td>0.040</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.066 (0.010)</td>
</tr>
<tr>
<td>Non-null setup #2</td>
<td>0.408</td>
<td>0.024</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.125 (0.020)</td>
</tr>
</tbody>
</table>

Fig. 4. Tail strength measure computed on some biological datasets. Shown are the TS measures along with 90% confidence intervals.

4. Examples

Figure 4 shows the tail strength measure and asymptotic 90% confidence intervals, applied to 12 different datasets. The datasets are summarized in Table 3. The first nine datasets are from microarray studies, and all report positive findings. Most of these are described in Dettling (2004), where some comparative analyses are also performed.
Tail strength measure for assessing univariate significance

Table 3. Summary of datasets for Figure 4

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>No. of samples</th>
<th>No. of features</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>FL</td>
<td>Microarray, survival</td>
<td>93</td>
<td>44,928</td>
<td>Dave et al. (2004)</td>
</tr>
<tr>
<td>Skin cancer</td>
<td>Microarray, two classes</td>
<td>58</td>
<td>12,625</td>
<td>Rieger et al. (2004)</td>
</tr>
<tr>
<td>Diffuse large cell lymphoma</td>
<td>Microarray, survival</td>
<td>240</td>
<td>7,399</td>
<td>Rosenwald et al. (2002)</td>
</tr>
<tr>
<td>Small round blue cell tumors</td>
<td>Microarray, four classes</td>
<td>63</td>
<td>2,308</td>
<td>Khan et al. (2001)</td>
</tr>
<tr>
<td>Colon cancer</td>
<td>Microarray, two classes</td>
<td>62</td>
<td>2,000</td>
<td>Alon et al. (1999)</td>
</tr>
<tr>
<td>Leukemia</td>
<td>Microarray, two classes</td>
<td>72</td>
<td>3,571</td>
<td>Golub et al. (1999)</td>
</tr>
<tr>
<td>Prostate cancer</td>
<td>Microarray, two classes</td>
<td>102</td>
<td>6,033</td>
<td>Singh et al. (2002)</td>
</tr>
<tr>
<td>Brain cancer</td>
<td>Microarray, five classes</td>
<td>22</td>
<td>5,597</td>
<td>Pomeroy et al. (2002)</td>
</tr>
<tr>
<td>Lymphoma</td>
<td>Microarray, two classes</td>
<td>62</td>
<td>4,026</td>
<td>Alizadeh et al. (2000)</td>
</tr>
<tr>
<td>aud-over</td>
<td>FMRI</td>
<td>15</td>
<td>187,782</td>
<td>Taylor and Worsley (2005)</td>
</tr>
<tr>
<td>aud-sent</td>
<td>FMRI</td>
<td>15</td>
<td>187,762</td>
<td>Taylor and Worsley (2005)</td>
</tr>
<tr>
<td>dtiTS</td>
<td>Diffusion tensor imaging</td>
<td>12</td>
<td>20,931</td>
<td>Schwartzmann et al. (2005)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>No. of samples</th>
<th>No. of features</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Tail strengths and misclassification rates (test set or cross-validated), for the classification problems in Table 3. Classification was done using nearest shrunken centroids

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Tail strength</th>
<th>Misclassification rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lymphoma</td>
<td>0.82</td>
<td>1.7</td>
</tr>
<tr>
<td>Brain</td>
<td>0.82</td>
<td>23.5</td>
</tr>
<tr>
<td>SRBCT</td>
<td>0.77</td>
<td>2.4</td>
</tr>
<tr>
<td>Prostate</td>
<td>0.70</td>
<td>8.9</td>
</tr>
<tr>
<td>Leukemia</td>
<td>0.68</td>
<td>3.6</td>
</tr>
<tr>
<td>Colon</td>
<td>0.55</td>
<td>13.5</td>
</tr>
<tr>
<td>Skin</td>
<td>0.18</td>
<td>20.1</td>
</tr>
</tbody>
</table>

SRBCT, small round blue cell tumors.

The remaining datasets are from neuroimaging studies. The datasets ‘aud-over’ and ‘aud-sent’ are from an auditory functional magnetic resonance imaging study (Taylor and Worsley, 2005) with aud-over being overall activation and aud-sent a measure of hemodynamic delay (Liao et al., 2002) in response to different sentences. The dataset ‘dtiTS’ comes from a 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 the datasets (except for FL) show significant (non-zero) tail strengths of various degrees. For the subset of classification problems among these studies, Table 4 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 4: the multiclass ‘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 are only 42 cases in this dataset, so that the training set on which the classifiers were trained had only 28 cases on an average. For the five classes, the classwise error rates were 15, 6, 6, 20 and 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 and 0.32. Hence, class 5 has both a high error rate and a lower tail strength. It seems that the overall tail strength, based on the $F$-statistic for all five classes, fails to capture the difficulty in predicting class five.

5. DISCUSSION

The tail strength measure is potentially useful for assessing the overall statistical significance of a set of hypothesis tests. It gives a quantitative measure of the overall strength of evidence against the global null hypothesis of ‘no association’ between a large set of features 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 degree of departure from the no association null 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 flurry of applications in genomics, there has been a resurgence of interest in this area (see, e.g. Dudoit et al., 2003, for a summary). Our work has a close relationship to the FDR approach to multiple testing, as we have shown in Section 2.2. There is a recent work of Efron (2005) (Section 5) in which quantities similar to tail strength are considered, based on a local version of the FDR.

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

$$\text{HC}_m = \max_{1 \leq i \leq a_0 m} \sqrt{m(i/m - p(i))} / \sqrt{p(i)(1 - p(i))},$$

for some $a_0 > 0$. This statistic is designed as an overall summary of the $p$-values, and they prove that it is optimal for detecting certain sparse patterns of $p$-values. They also show that the asymptotic $\alpha$ percentile for HC$_m$ is of the size $\sqrt{\log \log m}$. We attempted some numerical comparisons of HC$_m$ with tail strength on the datasets in this paper, but these were not successful. The presence of some very small $p$-values
Tail strength measure for assessing univariate significance

made the denominator very small and caused the statistic to get very large. In addition, it was not clear how to choose $\alpha_0$ 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 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, and editors and referees whose comments substantially improved this paper. Robert Tibshirani was partially supported by National Science Foundation grant DMS-9971405 and National Institutes of Health contract N01-HV-28183.

APPENDIX

A.1 Asymptotic properties of tail strength

In the 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, and give the proof of Theorem 2.1.

We can write

$$TS = -\frac{1}{m} \sum_{k=1}^{m} \left( p(k) - \frac{k}{m+1} \right) \frac{m+1}{k}$$

$$\simeq -\frac{1}{m} \sum_{k=1}^{m} \left( p(k) - \frac{k}{m} \right) \frac{m}{k}$$

$$= -\frac{1}{m} \sum_{k=1}^{m} \left( \frac{m}{k} \sum_{j=1}^{k} \left( p(j) - p(j-1) - \frac{1}{m} \right) \right)$$

$$= -\sum_{j=1}^{m} \left( p(j) - p(j-1) - \frac{1}{m} \right) \left( \sum_{k=j}^{m} \frac{1}{k} \right).$$

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 $\xi_i \sim \text{Exp}(1)$ 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 the 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 $\pi_0$ 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

$$F = \pi_0 \text{Unif}(0, 1) + (1 - \pi_0) F_1.$$
This is the mixture model used in the development of local FDR by Efron et al. (2001) and Efron and Tibshirani (2002). This assumption would not likely change the essence of our main result, only complicate the proofs.

We begin by expressing (2.1) in yet another way, in terms of quantile processes. Let

\[ \hat{Q}_m(q) = \hat{F}^{-1}_m(q) = \inf\{x: \hat{F}_m(x) > q\} \quad (A.1) \]

be the quantile process of the \( p \)-values and

\[ Q(q) = F^{-1}(q) = \inf\{x: F(x) > q\} \quad (A.2) \]

be the population quantile function.

Given the definition of \( \hat{Q}_m \), it is not hard to see that

\[ \hat{Q}_m\left(\frac{k}{m}\right) = p(k), \quad \forall \ 1 \leq k \leq m. \]

Using this fact, the expression (2.1) takes the form of a Riemann sum

\[ TS(p) \simeq \frac{1}{m} \sum_{k=1}^{m} \left( 1 - \frac{p(k)}{m} \right) \]

\[ = \frac{1}{m} \sum_{k=1}^{m} \left( \frac{k/m - Q(j/m)}{k/m} \right) + \frac{Q(j/m) - \hat{Q}_m(j/m)}{k/m} \]

\[ \xrightarrow{m \to \infty} \int_0^1 \left( 1 - \frac{Q(x)}{x} \right) dx + \lim_{m \to \infty} \int_0^1 \frac{Q(x) - \hat{Q}_m(x)}{x} dx. \quad (A.3) \]

Such an expression is simpler to analyze than (2.1), using some results from the theory of quantile processes.

If \( Q \) is Riemann integrable, then this 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 (2.8).

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

\[ m^{1/2}(\hat{Q}_m(x) - Q(x))_{0 \leq x \leq 1} \overset{D}{=} (B_x)_{0 \leq x \leq 1}, \quad (A.4) \]

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

\[ \text{Cov}(B(x), B(y)) = \min(x, y) - xy, \quad 0 \leq x, \ y \leq 1. \quad (A.5) \]

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)

\[ m^{1/2}(\hat{Q}_m(x) - Q(x))_{0 \leq x \leq 1} \Rightarrow \left( \frac{B_x}{f(Q(x))} \right)_{0 \leq x \leq 1}. \quad (A.6) \]
This suggests that
\[ m^{1/2} (TS - \mathbb{E}(TS)) \Rightarrow \int_0^1 \frac{B_x}{xf(Q(x))} \, dx \quad \text{as} \quad m \to \infty \Rightarrow \int_0^1 \frac{BF(x)}{F(x)} \, dx. \]

A straightforward application of Theorem 1 of Shorack (1972), combined with the comments above suffices to prove Theorem 2.1.

**Remark A.1** Actually, \( F \) need not even have a density, for the central limit to hold the above, though the expected value will be changed slightly. If \( F \) has density \( f \), then under the hypothesis \( F(x) \geq x \)
\[
\int_0^1 \int_0^1 \left( \frac{\min(F(x), F(y))}{F(x)F(y)} - 1 \right) \, dx \, dy = 2 \int_0^1 \int_y^1 \left( \frac{1}{F(x)} - 1 \right) \, dx \, dy \\
\leq 2 \int_y^1 \left( \frac{1}{x} - 1 \right) \, dx \, dy \\
= 1,
\]
so the variance under \( H_0 \) is an upper bound.

**References**


[Received August 25, 2005; revision November 29, 2005; accepted for publication December 1, 2005]