Order Thresholding and Goodness-of-Fit Testing

Michael Akritas

Joint work with
Ph.D. Student Min Hee Kim

Work supported in part by NSF grant
DMS-0805598.

Madison, 3 June 2010

Motivation

- Suppose $X_1, \ldots, X_n$ are independent with $X_i \sim N(\mu_i, 1)$, and we want to test $H_0 : \mu_1 = \cdots = \mu_n = 0$ *vs* $H_a : H_0$ is not true.
- The obvious test statistic is

$$W_n = \sum_{i=1}^{n} Y_i, \quad \text{where} \quad Y_i = X_i^2.$$

- The asymptotic, as $n \to \infty$, distribution of this statistic under alternatives $\boldsymbol{\mu}$ which tend to infinity at rates $||\boldsymbol{\mu}||^2 = o(\sqrt{n})$ is the same as under the null.
- The use of thresholding methods in hypothesis testing is motivated by attempts to improve the power of the chi-square statistic.

Signal-to-Noise Ratio

- Alternatives $\boldsymbol{\mu}$ such that $\mu_i = 0$ for "most" i will be called *low signal-to-noise ratio* alternatives.
- For such alternatives there is an intuitive explanation for the low power of the chi-square statistic:

  Too much "noise" drowns out the few "signals"

  where "noise" refers to the central χ_1^2 r.v.'s, and "signals" refers to the non-central χ_1^2 r.v.'s.
- The basic idea of thresholding methods is to somehow eliminate the noise and focus on the signal.

Work supported in part by NSF grant DMS-0805598.

Hard and Order Thresholding

- The HT statistic is

$$T_H(\delta_n) = \sum_{j=1}^{n} Y_j I(Y_j > \delta_n), \quad \delta_n = 2\log(n \log^{-2} n).$$

- The OT statistic is

$$T_O(k_n) = \sum_{i=1}^{n} I(i > n - k_n) Y_{i,n}, \quad k_n < n, \; k_n \to \infty.$$

where $Y_{1,n} < \cdots < Y_{n,n}$ are the ordered Y_is.

Hard and Order Thresholding

- The HT statistic is

$$T_H(\delta_n) = \sum_{j=1}^{n} Y_j I(Y_j > \delta_n), \quad \delta_n = 2\log(n \log^{-2} n).$$

- The OT statistic is

$$T_O(k_n) = \sum_{i=1}^{n} I(i > n - k_n) Y_{i,n}, \quad k_n < n, \; k_n \to \infty.$$

  where $Y_{1,n} < \cdots < Y_{n,n}$ are the ordered Y_is.
- Their form makes it clear that they eliminate excess "noise" by focusing on the largest squared values. But they differ in the way they go about doing it.

	Empirical Power $n = 500$		
	$T_H(5.122)$	$T_O(\widehat{k}^{opt}_{500})$	W_n
H_{10}	0.817	0.908	0.744
H_{12}	0.783	0.903	0.709
H_{14}	0.734	0.864	0.649
H_{16}	0.564	0.707	0.484
H_{17}	0.529	0.675	0.432
H_{18}	0.435	0.584	0.373
H_{19}	0.402	0.570	0.347
H_{20}	0.380	0.547	0.308
H_{21}	0.390	0.555	0.319
H_{22}	0.364	0.534	0.281
H_{23}	0.362	0.517	0.279

How about high signal-to-noise ratio?

- In high signal-to-noise ratio situations, i.e. if all $\mu_i = c$, the threshold statistics will not improve the power ...

How about high signal-to-noise ratio?

- In high signal-to-noise ratio situations, i.e. if all $\mu_i = c$, the threshold statistics will not improve the power ... unless we can transform the data so the signal IS concentrated in a few locations.

How about high signal-to-noise ratio?

- In high signal-to-noise ratio situations, i.e. if all $\mu_i = c$, the threshold statistics will not improve the power ... unless we can transform the data so the signal IS concentrated in a few locations. Here is the basic idea:
- Set $\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$, for the data vector, let $\boldsymbol{\Gamma}$ be an orthonormal matrix, and transform $\mathbf{X}$ to

$$\boldsymbol{\Gamma}\mathbf{X} = \boldsymbol{\Gamma}\mu + \boldsymbol{\Gamma}\epsilon \text{ or } \mathbf{X}_{\boldsymbol{\Gamma}} = \boldsymbol{\mu}_{\boldsymbol{\Gamma}} + \boldsymbol{\epsilon}_{\boldsymbol{\Gamma}}.$$

so that $||\boldsymbol{\mu}|| = ||\boldsymbol{\mu}_{\boldsymbol{\Gamma}}||$ and $\boldsymbol{\epsilon}_{\boldsymbol{\Gamma}} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$

How about high signal-to-noise ratio?

- In high signal-to-noise ratio situations, i.e. if all $\mu_i = c$, the threshold statistics will not improve the power ... unless we can transform the data so the signal IS concentrated in a few locations. Here is the basic idea:
- Set $\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$, for the data vector, let $\boldsymbol{\Gamma}$ be an orthonormal matrix, and transform $\mathbf{X}$ to

$$\boldsymbol{\Gamma}\mathbf{X} = \boldsymbol{\Gamma}\mu + \boldsymbol{\Gamma}\epsilon \quad \text{or} \quad \mathbf{X}_{\boldsymbol{\Gamma}} = \mu_{\boldsymbol{\Gamma}} + \epsilon_{\boldsymbol{\Gamma}}.$$

  so that $||\boldsymbol{\mu}|| = ||\boldsymbol{\mu}_{\boldsymbol{\Gamma}}||$ and $\epsilon_{\boldsymbol{\Gamma}} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$
- W_n has the same power on $\mathbf{X}$ as it does on $\mathbf{X}_{\boldsymbol{\Gamma}}$.

How about high signal-to-noise ratio?

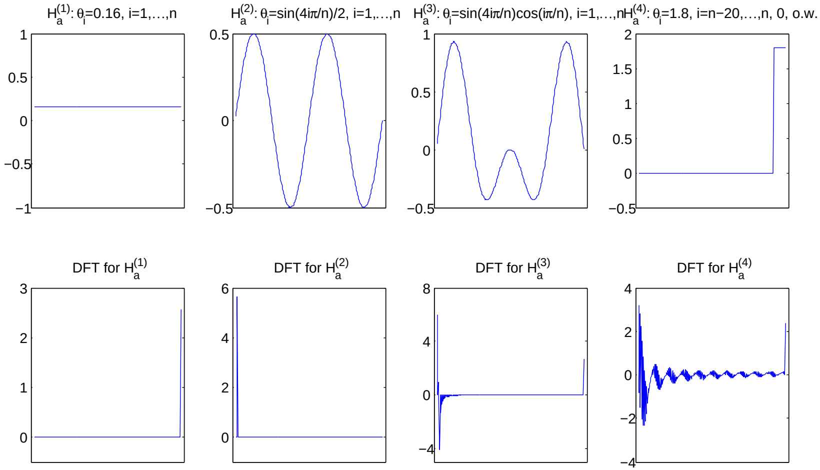
- In high signal-to-noise ratio situations, i.e. if all $\mu_i = c$, the threshold statistics will not improve the power ... unless we can transform the data so the signal IS concentrated in a few locations. Here is the basic idea:
- Set $\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$, for the data vector, let $\boldsymbol{\Gamma}$ be an orthonormal matrix, and transform $\mathbf{X}$ to

$$\boldsymbol{\Gamma}\mathbf{X} = \boldsymbol{\Gamma}\mu + \boldsymbol{\Gamma}\epsilon \quad \text{or} \quad \mathbf{X}_{\boldsymbol{\Gamma}} = \mu_{\boldsymbol{\Gamma}} + \epsilon_{\boldsymbol{\Gamma}}.$$

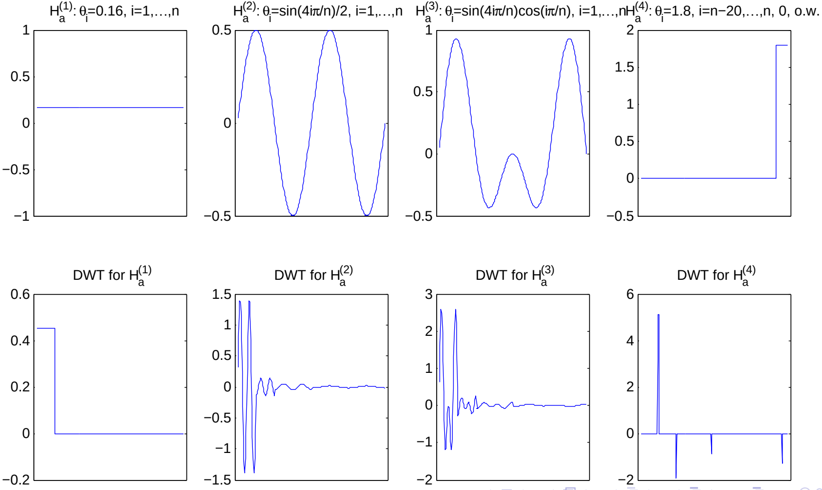
so that $||\boldsymbol{\mu}|| = ||\boldsymbol{\mu}_{\boldsymbol{\Gamma}}||$ and $\boldsymbol{\epsilon}_{\boldsymbol{\Gamma}} \sim N_n(\mathbf{0}, \sigma^2\mathbf{I})$

- W_n has the same power on $\mathbf{X}$ as it does on $\mathbf{X}_{\boldsymbol{\Gamma}}$. Thus, if $\boldsymbol{\mu}_{\boldsymbol{\Gamma}}$ is a low signal-to-noise ratio alternative, thresholding will improve the power.

The Discrete Fourier Transformation (DFT)



The Discrete Wavelet Transformation (DWT)



	Before DFT		After DFT	
	$\sum_{i=1}^{255} X_i^2$	$T_O(\widehat{k}_{255}^{opt})$	$\sum_{i=1}^{255} X_{\boldsymbol{\Gamma},i}^2$	$T_O(\widehat{k}_{255}^{opt})$
H_0	0.051	0.056	0.051	0.075
$H_a^{(1)}$	0.091	0.087	0.091	0.148
$H_a^{(2)}$	0.397	0.294	0.397	0.841
$H_a^{(3)}$	0.816	0.671	0.816	0.992
$H_a^{(4)}$	0.850	0.877	0.850	0.925

	Before DWT		After DWT	
	$\sum_{i=1}^{256} X_i^2$	$T_O(\widehat{k}_{256}^{opt})$	$\sum_{i=1}^{256} X_{\Gamma,i}^2$	$T_O(\widehat{k}_{256}^{opt})$
H_0	0.051	0.062	0.051	0.058
$H_a^{(1)}$	0.093	0.094	0.093	0.096
$H_a^{(2)}$	0.396	0.313	0.396	0.400
$H_a^{(3)}$	0.813	0.661	0.813	0.881
$H_a^{(4)}$	0.861	0.880	0.861	0.998

- $X_1, \ldots, X_n$ iid G. Wish to test $H_0 : G = G_0$. Transforming to $U_i = G_0(X_i)$ we consider testing $H_0 : G(t) = t,\ \forall 0 < t < 1$.
- Karl Pearson's 1900 χ^2 test uses a partition of n_{bin} cells and tests for the parameters of a multinomial distribution.
- As such it has been widely applied also in contingency tables.
- Mann and Wald (1942) were the first to establish the power advantages of letting n_{bin} tend to infinity with n, and found $n_{bin} = n^{2/5}$ to be the optimal rate.
- For a corresponding development in the area of contingency tables see Holst (1972), Morris (1975), and Koehler and Larntz (1980).
- We will explore the use of thresholding methods to further improve the power.

- Partition the interval $(0,1]$ into the n_{bin} subintervals $J_i = ((i-1)/n_{bin}, i/n_{bin}]$, $i = 1, \ldots, n_{bin}$, and set

$$N_j = n \int_{J_j} \mathrm{d}\widehat{G}_n(x), \text{ so that } S_j = \frac{N_j - n/n_{bin}}{\sqrt{n(n_{bin}-1)/n_{bin}^2}} \;\dot{\sim}\; N(0,1).$$

- Transform $\mathbf{S} = (S_1, \ldots, S_{n_{bin}})'$ by either the DFT or DWT:

$$\mathbf{S}_{\mathbf{\Gamma}} = (S_{\mathbf{\Gamma},1}, \ldots, S_{\mathbf{\Gamma},n_{bin}})'$$

and use the OT statistic on $\mathbf{S}_{\mathbf{\Gamma}}$:

$$T_O(k_{n_{bin}}) = \sum_{j=n_{bin}-k_{n_{bin}}+1}^{n_{bin}} S^2_{\mathbf{\Gamma},(j)}.$$

- The S_j's are not quite normal and are not quite independent. Thus, application of the OT methodology needs justification. The asymptotic justification consists of two parts:
  1. Use a strong approximation of the empirical process by a Brownian bridge, to represent each S_j as:

     $$S_j = X_j + V_j, \quad j = 1, \ldots, n_{bin},$$

     where $X_j \sim$ iid $N(0,1)$ and $V_j = -n_{bin}^{-1/2} W_n(1) + R_n$.
  2. Account for the fact that the ordering of the S_j^2's is not exactly the same as the ordering of the $Y_j = X_j^2$'s.

- The basic idea for accounting for the lack of ordering of the $Y_{(j)}$'s is to create a "buffer zone" which separates "most" of the largest $k_{n_{bin}}$ order statistics that we want to include in the OT statistic (i.e. $Y_{n_{bin}-k_{n_{bin}}+1,n_{bin}}, \ldots, Y_{n_{bin},n_{bin}}$) from "most" of the smallest ones that we want to exclude.
- We now make this idea precise.

Order Work supported in part by NSF grant DMS-0805598.

It is shown that the sequence of integers $a_{n_{bin}} = k_{n_{bin}} + k_{n_{bin}}^{1/2-\delta}$ satisfies

(1) $k_{n_{bin}} < a_{n_{bin}} \leq n_{bin}$

(2) $$\frac{\sum_{j=n_{bin}-a_{n_{bin}}+1}^{n_{bin}} Y_{j,n_{bin}} - n_{bin}\mu_{n_{bin}}(k_{n_{bin}})}{\sqrt{n_{bin}}\sigma_{n_{bin}}(k_{n_{bin}})} \xrightarrow{d} N(0,1),$$
as $n_{bin} \rightarrow \infty$

(3) $$P\left(\left\{Y_{(n_{bin}-k_{n_{bin}}+1)}, \ldots, Y_{(n_{bin})}\right\} \subset \left\{Y_{n_{bin}-a_{n_{bin}}+1,n_{bin}}, \ldots, Y_{n_{bin},n_{bin}}\right\}\right) \rightarrow 1.$$

Example 1. Mixture model alternatives of the form:

$$H_0 : G = N(0,1) \text{ versus } H_a : G = 0.7N\left(\mu/0.7, 1\right) + 0.3N\left(-\mu/0.3, 1\right),$$

				After DFT		After DWT	
μ	Pearson	KS	CVM	$T_H(1.93)$	$T_O(6)$	$T_H(1.96)$	$T_O(6)$
.400	.9668	.8344	.8967	.9273	.9804	.9410	.9637
.367	.8777	.6692	.7393	.8206	.9124	.8334	.8777
.333	.7161	.4881	.5413	.6553	.7724	.6795	.7161
.300	.5087	.3244	.3473	.4787	.5675	.4951	.5110
.267	.3184	.2132	.2150	.3322	.3711	.3594	.3276
.233	.1940	.1480	.1396	.2346	.2194	.2526	.2025
.200	.1059	.0935	.0891	.1579	.1215	.1830	.1180

Example 2. Beta alternatives of the form: $H_0 : G = Uniform(0,1)$ versus $H_a : G = Beta(1.1, c)$.

				After DFT		After DWT	
c	Pearson	KS	CVM	$T_H(1.93)$	$T_O(6)$	$T_H(1.96)$	$T_O(6)$
1.8	.9842	1	1	.9987	.9926	.9983	.9949
1.7	.9391	.9999	1	.9885	.9595	.9890	.9703
1.6	.8171	.9972	.9988	.9437	.8550	.9467	.8789
1.5	.6126	.9720	.9853	.8187	.6631	.8289	.6809
1.4	.3871	.8505	.8880	.5939	.4285	.6247	.4284
1.3	.2045	.5330	.5841	.3597	.2324	.3909	.2183
1.2	.1169	.1964	.2110	.2012	.1329	.2195	.1158
1.1	.0687	.0552	.0532	.1332	.0837	.1542	.0771
1.0	.0741	.1697	.1957	.1406	.0850	.1586	.0806
0.9	.2113	.5684	.6523	.2806	.2215	.3123	.2214
0.8	.6243	.9303	.9609	.6083	.5956	.6455	.6732